Chapter 3

English-American Speech Relationships: A Quantitative Approach

Abstract. This study applies quantitative techniques – measures of linguistic distance, cluster analysis, principal component analysis, and regression analysis – to data on English speech variants in England and America drawn mainly from Kurath and McDavid's (1961) *The Pronunciation of English in the Atlantic States*. The analysis yields measures of similarity among English and American speakers, distinguishes clusters of speakers with similar speech patterns, and isolates groups of variants that distinguish those groups of speakers. The results are consistent with a model of new-dialect formation in the American colonies, involving competition within and selection from a pool of variants introduced by speakers from different dialect regions. The patterns of similarity appear to be largely consistent with the historical evidence of migrations from seventeenth- and eighteenth-century Britain to North America, lending support to the hypothesis of regional English origins for some important differences in American dialects, and suggesting mainly southeastern English influence on American speech, with somewhat greater southeastern influence on New England speech and southwestern influence in the American south.[1]

[1]The final, definitive version of this paper has been published in the *Journal of English Linguistics*, Vol. 33, No. 2, (June 2005) by SAGE Publications, Inc. On-line version available at `http://eng.sagepub.com/cgi/content/abstract/33/2/99`.

3.1 Introduction

Linguists and layfolk alike have devoted much thought to the origins of American English dialect forms. Some have emphasized the importance of non-European and non-English influences on the development of American speech. Others stress continuities with traditional English forms from various parts of the British Isles, even arguing that "some of today's most noticeable dialect differences can be traced directly back to the British dialects of the seventeenth and eighteenth centuries."[2] The latter view underlies such efforts as that of Cleanth Brooks (1935), who undertook a detailed review of the British dialect material in Joseph Wright's *English Dialect Dictionary* (1898-1905), systematically comparing over 100 forms found in the speech of his native region of Alabama (or used by characters in southern American literary works) with forms found in different parts of England. Brooks concluded that southern American speech forms were derived primarily from earlier dialects spoken mainly in southern England, particularly in southwestern England.

An important development in dialect research occurred three years later when, in conjunction with his research for the *Linguistic Atlas of New England* (*LANE*) and the *Linguistic Atlas of the Middle and Southern Atlantic States* (*LAMSAS*), Guy Lowman conducted a wide-meshed survey of rural speech in southern England to permit more informed comparisons between the speech forms of England and America than had been possible previously. After Lowman's untimely death in 1941, the materials from his English survey were interpreted and presented by others. Viereck (1975) presented extensive lexical and grammatical results as well as much of the supporting methodological detail, and Kurath and Lowman (1970) – *The Dialect Structure of Southern English* – summarized the phonological material, presented some tentative conclusions about the structure of southern English dialects, and noted some correspondences between southern English forms and those found in different regions of the United States.

Kurath and McDavid (1961) – *The Pronunciation of English in the Atlantic States* – published even more phonological detail from Lowman's English research, presenting results from *LANE*, *LAMSAS*, and his survey in a series of annotated maps illustrating the occurrence of different variants of English phonemes in the Atlantic states as well as in southern England. The maps reveal a great variety of forms in both England and America during the first half of the twentieth century. Without presenting any systematic quantitative assessment or comparisons, moreover, the maps clearly illustrate that most variants found in use by American speakers could also be found in use by

[2] For a relatively Anglocentric view, see Wolfram and Schilling-Estes (1998), p. 93. For a quite contrary view, see Dillard (1992), Chapter 1.

traditional southern English speakers, although American speech was considerably less variable than that of rural southern England.

During the past two generations, researchers have continued to uncover sources of American speech forms from southern England and elsewhere. Recently, for example, Montgomery (2001) traced the influence – predominantly on vocabulary – of eighteenth-century Scots-Irish immigrants on speech in the Appalachians and Upper South. Wright (2003) uncovered a variety of grammatical features associated with Southern American English in prisoners' narratives from early seventeenth-century London. Algeo (2003), discussing the origins of Southern American English, argues for "multiple lines of descent" from southern and western English, Scotch-Irish, African, and other influences. Orton and Dieth (1962) have also improved on Lowman's English research by completing and publishing a *Survey of English Dialects*, covering all of England.

Historians, too, have contributed to linguistic research by tracing differences in the British origins of settlers of different regions of North America. Notable examples are Bailyn (1986) and Fischer (1989), who document extensive processes of internal migration from all over the British Isles to London and a predominance of emigration to most of the colonies from London and surrounding regions, but also show somewhat greater-than-average migration from East Anglia to New England, from the Midlands to Pennsylvania, from the West Country to Virginia, and from Scotland and northern Ireland to the backcountry.

A related strand of research – one that benefits from the more recent nature of the phenomena in question and, consequently, greater availability of data – focuses on the development of English dialects in other, more recently settled colonies such as Australia and New Zealand. Trudgill (1986) compared a number of phonetic characteristics of Australian English with those of English dialects, noting a very close relationship between Australian English and the speech forms of London and Essex, and concluded that Australian English is "a mixed dialect which grew up in Australia out of the interaction of south-eastern English forms with East Anglian, Irish, Scottish and other dialects."[3] More recently, Trudgill (2004) has closely studied the process of new-dialect development following contact among speakers of different dialects of English, noting that such contact "would have led to the appearance of new, mixed dialects not precisely like any dialect spoken in the homeland."[4] In the case of New Zealand, Trudgill tracks the process of dialect formation from a first stage of dialect contact among immigrants from a variety of British origins, through a second stage in which first-generation speakers choose from the variety of

[3] Trudgill (1986), p. 142.
[4] Trudgill (2004), p. 2.

speech forms available to them, to a third stage in which a relatively uniform dialect emerges among second-generation speakers.

From such strands of research, many dialectologists conclude that differences in migration patterns and settlement histories are likely to have contributed to significant differences among American regional dialects, with a largely but not exclusively English influence. The processes that led to such differentiation were presumably as complex as those documented by Trudgill for New Zealand. They were likely driven in part by what Mufwene (1996) has called the *founder effect*, by which the speech forms of the earliest settlers have an inherent advantage in the process of survival and propagation, analogous to the biological advantage of their genes. In addition, variants most frequently used by the largest group of settlers were probably more likely to dominate as speakers adapted their speech habits to those of their most frequent interlocutors. Some variants may have acquired higher prestige and spread; others may have been stigmatized and therefore declined. Speakers may have come to associate particular variants with ethnic or regional identities, tying the fate of those variants with that of the identities. Kretzschmar (2002) emphasizes that the processes were largely local, noting that despite the development of American forms of English that appeared remarkably uniform to many British observers, records of colonial speech patterns reveal a great deal of regional and even local diversity, belying any simple narrative involving the emergence of regional dialects from a melting, mixing, or weaving of forms brought during settlement.

Until recently, many experts had concluded not only that differences among American regional dialects were largely the result of settlement processes, but that many if not most regional differences in the Atlantic States were largely formed by the American Revolution. More recent research has emphasized the importance of innovations that occurred after the period of colonization and settlement. Schneider (2003), examining Orton et al. (1978) for possible English sources of 25 pronunciation features common in Southern American English, has found 11 in southwestern England and eight in the southeast (with considerable overlap), but only four to five in other regions. Schneider concludes that there is some "limited continuity of forms derived from British dialects" but "also a great deal of internal dynamics to be observed ... and ... strong evidence for much innovation."[5] Similarly, linguists such as Bailey (1997), while accepting that features of colonial and early post-colonial varieties were likely largely a consequence of settlement history, have shown that some common Southern American English features (such as the *pin*/*pen* merger) that may have been in sporadic use not long after the Revolution did not become common until long after the period of settlement.

[5]Schneider (2003), p. 34.

In the meantime, linguists have made significant progress in a complementary strand of research, the development of methods of quantifying differences among speech forms.[6] Moving well beyond the isogloss methods characteristic of earlier work, that research has provided a variety of methods of measuring distances between sound segments, either measured acoustically or, as in the case of the Lowman data, impressionistically, as well as methods to measure distances between speakers or groups of speakers, based on aggregates of measures between specific sound segments. Dialectologists have employed such tools to great effect, as for example in Heeringa's (2004) analysis of Dutch and Norwegian dialects or Nerbonne's (forthcoming) examination of American speech in Virginia and North Carolina.

To date, however, no such dialectometric analysis has been applied to a data set including both English and American speakers. An effort to quantify linguistic distances among English and American speech forms and speakers recorded in the twentieth century may provide some insights into how varieties of American English have developed over time, and perhaps even how English variants were selected in the development of new dialects in the American colonies. Such an effort is hampered, of course, by temporal distance: the fact that the earliest English settlers arrived nearly four centuries ago raises serious questions as to whether a synchronic comparison of twentieth-century American and English forms can provide any insights at all into dialect developments during colonization. However, the barrier may not be quite as profound as it seems at first sight: parts of the Atlantic coast were still being settled at the end of the seventeenth century, and some of the speakers interviewed by Lowman were born in the middle of the nineteenth century. In at least some cases, therefore, the time distance is more like 150 years rather than 400 years – enough distance to raise questions as to whether the proposed comparisons can yield historical insights, certainly, but not enough distance to preclude the possibility altogether.

This study discusses characteristics of the data presented in Kurath and McDavid (1961) and Kurath and Lowman (1970) discussed above, and applies quantitative techniques to that data to characterize the degrees of and patterns of similarity among a subset of American and English speakers, to distinguish clusters of speakers with similar speech patterns, and to isolate groups of variants that distinguish those groups of speakers. The results provide insights into the differences and similarities among dialects, and can be used to make tentative inferences about the processes of new-dialect formation that might have occurred in the development of American regional dialects.

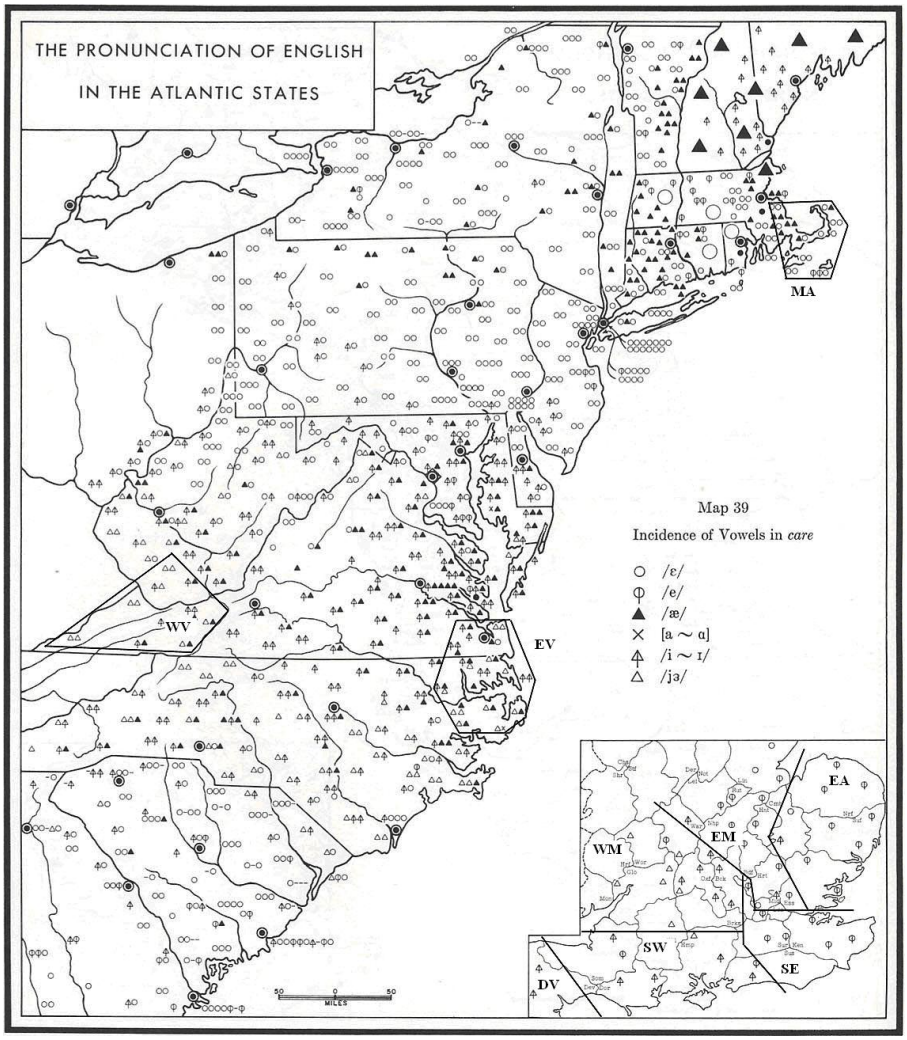
[6] Heeringa (2004) provides a very useful summary of much of that research.

3.2 Assembling Data from *The Pronunciation of English in the Atlantic States* and *The Dialect Structure of Southern England*

In an ideal world, this analysis would draw on easily accessible and interpretable data, preferably collected by one person using a uniform methodology, describing as many speech forms as possible from informants from different regions. The real data that best (though quite imperfectly) meet those criteria appear to be those presented in the maps of two works: Kurath and McDavid (1961) – *The Pronunciation of English in the Atlantic States* (henceforth *PEAS*) – and Kurath and Lowman (1970) – *The Dialect Structure of Southern England* (*DSSE*). Figure 3.1, taken from *PEAS*, shows a sample of that data – a map of the occurrence of six different vocalic variants used in *care*, each variant distinguished by a different type of marker. Each marker represents the pronunciation of a specific informant from a given location, although there are often two or more informants from a location, occasionally more than one variant per informant, and occasionally no data for an informant. Kurath and McDavid sometimes distinguished regions in which a particular variant was widespread by using a large marker (the large black triangles in New England, for instance), indicating the occurrence of other variants with regularly-sized markers. It is usually rather easy to associate a marker with an informant described in Kurath (1939) – the *LANE* handbook – or Kretzschmar et al. (1994) – the *LAMSAS* handbook – or Viereck (1975), which describes Lowman's English informants. The markers, however, were not always placed in exactly the same place on each map and the interpretive process occasionally becomes somewhat creative.

Altogether, 84 maps in *PEAS* provide information about 275 phonetic variants recorded in 76 different words (or, in one case, a phrase) in England and America. Two of the maps also permit us to distinguish between informants who show a single pronunciation for the vowel in words pronounced with [**aː**] or [**ai**] in London standard Middle English and those who do not. In addition, six maps of Lowman's English data in *DSSE* provide data that can be used to tabulate the southern English usage of variants in one or more words or to distinguish a merger (Middle English [**ou**] and [**ɔː**]). For one of those words, maps from *PEAS* provide the American usages; for the others, Americans universally have a single, obvious usage (for example, unvoiced fricatives in words like *furrow*, *fog*, and *frost*). Altogether, the maps in *PEAS* and *DSSE* make it possible to distinguish and tabulate the English and American informants' usage of 284 variants in 81 different words – many of which are particularly notorious for a variety of nonstandard pronunciations – and the presence or absence of two mergers. The words cover nearly all phonemes in standard

Figure 3.1: Location of English and American Informants and Regions



Reprinted with permission from *The Pronunciation of English in the Atlantic States*, by Hans Kurath and Raven I, McDavid, Jr., University of Michigan Press, 1961.

British English and American English – unstressed, short, and long vowels, diphthongs (including many rhotic ones), and a number of consonants – usually with one to four words for a particular phoneme but with five or more words for a few. The full list – 83 cases involving a total of 288 variants (including the mergers), as shown in Table B.1 in Appendix B – constitute a fairly wide if not fully comprehensive tabulation of phonetic variation in southern English and American speech.

A few qualifications are in order. First, in considering the utility of this data set for comparing speech forms, it is important to keep in mind that the detailed responses recorded by the interviewers were grouped into variants or "allophones" by Kurath and McDavid, causing a significant loss of real diversity in the characterizations used here. In this sense, some of the variability in the data has already been eliminated by Kurath and McDavid's choices of how to classify responses into variants. The choices reflect those researchers' views of the structure of English dialects, and as a consequence their views may well be reflected in any results derived from analysis of the data. Second, in several cases the nature of the data requires us to create a residual variant that in fact constitutes a group of variants that cannot be readily distinguished. In those cases, too, the data (and any analysis of it) tends to understate the actual variability of the speech forms. Third, in a few cases, the representation of the data on the maps makes it very difficult to determine which of two or three possible informants in a given locality gave the observation; in these cases, the attribution is made arbitrarily. All three of those limitations could be overcome by future research drawing from the interviewers' original records.

A further qualification is that in a handful of cases, mainly in maps from *DSSE*, the maps present data that is more in the nature of a frequency (e.g., the presence of a variant in one or more of six words) than nominal data indicating simple presence or absence of a variant in a single word. As a general rule it is inadvisable to mix nominal data and frequency data. In the case of the Lowman data, however, it does not appear to be a gross violation of that rule to interpret the nominal data as 0 percent or 100 percent frequency usage of a particular variant in a specific context by a specific informant, and to combine it with data that measures 0 percent to 100 percent frequency of a particular variant in several contexts by the same informant.

Because it is focused primarily on the transmission of speech forms during the early settlement of the English colonies, this analysis draws from the maps to obtain records only for informants from England and from three relatively restricted areas in the United States. (Obviously, a great deal more insight may be gleaned by expanding the analysis to include records from more of the American informants presented in *PEAS*.) The analysis includes records for informants in two American regions that were settled extremely early – 22 informants in a region surrounding Plymouth, Massachusetts and 31 infor-

mants in a region along the southern Virginia and northern North Carolina coast. In addition, the analysis includes records for 19 informants from a region encompassing southernmost West Virginia and southwestern Virginia – the geographic center of Carver's (1987) Upper South dialect region. That latter choice reflects the author's particular interest in understanding the origins of Appalachian speech and in testing the popular perception that Appalachian speakers, being particularly isolated, retain archaic speech forms to a greater degree than do speakers of many other dialects.

Table B.2 in Appendix B presents information describing the informants and their interviews.[7] Guy Lowman interviewed most but, unfortunately, not all of the informants: seven informants in southeastern England were interviewed by Henry Collins, 13 of the informants in Massachusetts were interviewed by Cassil Reynard, and two more were interviewed by Miles Hanley. Most of the interviews were conducted between 1934 and 1940, with the exception of Collins' which were conducted in 1950. All but two of the English informants whose gender can be identified from the records were male, so chosen, presumably, on the principle that men tend to retain more old-fashioned speech forms. All of the English informants could be classified as older "folk" speakers of traditional rural dialects – that is, speakers with "local usage subject to a minimum of education and other outside influence."[8] In most cases we know only that they were generally over the age of 60, and therefore typically born before 1878.

The directors of the *Linguistic Atlas* projects attempted to secure a representative cross-section of regional speech forms acquired mainly during the second half of the nineteenth century, with moderate but not exclusive emphasis on folk speech. In the regions examined in this study, all of the American informants were white; nearly all lived in rural settings or in small towns and came from families that were long-established in the region. All but one of the Massachusetts informants – but only 26 of the 50 southern American informants – were male. The interviewers classified more than half (39, or 54 percent) of the American informants as folk speakers. They classified 27 (38 percent) as "common" speakers with "local usage subject to a moderate amount of education ... private reading, and other external contacts" and the remaining six (8 percent) as "cultivated" speakers with "wide reading and elevated local cultural traditions, generally but not always with higher education." The typical American speaker was born in 1872 and was 64 years old at the time of the interview. The youngest was born in 1918 and was 18; and the eldest was born in 1846 and was 88.

[7] The descriptions of English informants are taken from Viereck (1975); of Massachusetts informants, from Kurath (1939), and of southern American informants, from Kretzschmar et al. (1994).

[8] For these and the following definitions, see Kretzschmar et al. (1994), p. 25.

To summarize, data for each of the 59 English and 72 American informants are represented by a vector of 288 variables. Most take a value of 1 or 0, indicating the informant's use or lack of use of a particular variant. In a handful of cases, the value indicates the frequency with which the informant uses the variant in a set of words. The interviewers occasionally elicited more than one variant per informant, in which case both are represented in the data set. Less than two percent of the data is missing, where interviewers did not ask a question or did not get a useful reply. More than a third of the observations are missing no data and only a handful is missing more than five percent. The data set is available as a Microsoft Excel spreadsheet on request from the author.

3.3 Choice of Quantitative Techniques

A variety of quantitative techniques can be brought to bear to analyze the variation in the sample and the degree of similarity among speakers within and among regions, to distinguish groups of speakers with similar speech patterns, and to distinguish groups of variants that characterize those groups' speech patterns. Many such methods have already been applied in studies of dialect geography but not, to my knowledge, to phonetic data from *PEAS* and *DSSE*. In this study, I present the analysis in the following sequence:

- **Using cluster analysis to find dialect regions**. A variety of clustering techniques can be used to group informants on the basis of some measurement of similarity of their speech patterns. Ideally, the groups will be interpretable as geographically contiguous dialect regions. By distinguishing a reasonably coherent set of dialect regions, the cluster analysis lays the basis for examining the geographic distribution of variants and for measuring degrees of similarity among speakers from different regions.

- **Analyzing the distribution of variants.** A great deal can be learned by simply examining how variants are distributed among speakers in different regions – how many (and which) variants appear in different regions, and how many are shared between and among regions.

- **Applying measures of similarity (distance measures) among informants.** Even more information can be uncovered by measuring degrees of similarity between and among individual speakers within and among regions. By helping to distinguish degrees of difference among varieties of speech, distance measures provide a reasonably objective gauge of whether (and which) English and American informants' speech forms

are dramatically different or relatively similar. A number of measures are available, including the percentage of a speaker's total number of variants that he or she shares in common with other speakers; Pearson correlations, Euclidean distances, or cosines between vectors of values of variants; and various measures of linguistic distance or genetic distance. This analysis focuses on the simplest – measures of the percentage of shared variants – and on a quantitative measure of the articulatory or acoustic differences between speakers' variants; that is, a measure of linguistic distance, which is arguably most appropriate to the data. (The cluster analysis discussed above and the principal component analysis discussed below rely on other measures, discussed in the following sections.)

- **Principal component analysis.** With data sets that measure variation along a large number of dimensions (such as the occurrence, nonoccurrence, or frequency of occurrence of variant pronunciations), principal component analysis can be used to reduce the information to a smaller number of dimensions that, ideally, have clear interpretations. In this case, principal component analysis can be used to determine whether (and how strongly) groups of variants tend to occur together and whether (and what) groups of speakers use those groups of variants together. Ideally, the principal components that are the output of such an analysis will be interpretable as linguistically relevant groups of variant pronunciations, and will have clear interpretations in terms of dialect geography.

- **Multiple regression analysis.** Finally, regression analysis can be used to test for relationships among variables, and may provide insights into geographical characteristics of the distribution of variants. In this case, I use regressions to test for a statistically significant relationship between the degree of similarity between English and American speakers and the proximity of the English speakers to London. Such a relationship, if it exists, may provide support for the hypothesis that speech in or near the London metropolitan area played a key role in the development in American speech varieties.

The techniques used in this study are implemented using the Statistical Package for the Social Sciences (SPSS) for Windows Version 7.5 or a Fortran-based program written by and available on request from the author.

3.4 Using Cluster Analysis to Distinguish Dialect Regions

Cluster analysis refers to a large set of mathematical procedures that divide data into classes based on relationships within the data, thus dramatically reducing variation along a number of dimensions in the data set to a single set of clusters.[9] In this study, clustering methods are used to classify informants whose speech is similar according to some quantitative measure into distinct groups.

Clustering techniques include *non-hierarchical* methods, in which the data is divided into an arbitrary number of classes and each observation is assigned to a particular class, and *hierarchical* methods, in which classes may be divided into subclasses. Non-hierarchical methods exclude any relation among clusters, while hierarchical methods allow subclusters to be more or less closely related as members of larger clusters, and a given observation may a member of a several subclusters, e.g., large cluster of clusters, one of that group's subclusters, and so on (hence the notion of hierarchy). Hierarchical methods include divisive techniques, which divide and subdivide a data set into subsets until some predetermined limit is reached, and agglomerative methods, which start with each observation as a separate cluster, join the most similar ones, and continue to join the resulting clusters until all clusters have been united.

Every clustering method and distance measure has strengths and weaknesses, depending on the actual distribution and type of the data. This analysis tries to compensate for the weaknesses of different methods and measures by using several of each. First, it applies a non-hierarchical clustering method and a Euclidean distance measure to explore how the speakers cluster as the number of clusters is increased from two to ten. Second, it applies twelve different hierarchical analyses, using four different agglomerative methods – single linkage (nearest neighbor), complete linkage (furthest neighbor), average linkage between groups, and average linkage within groups – and three different distance measures – Euclidean distances, Pearson correlations, and cosines.[10] The multiplicity of approaches helps provide insights into the robustness of the clusters produced under different approaches.

[9] There are many clustering procedures, some of which involve the use statistical probabilities or statistical measures, but in general the procedures are not properly thought of as statistical since most do not assess the probability that observations are "correctly" classified.

[10] In principle, clustering techniques can incorporate any distance (or similarity) measure one wishes, with some measures being more appropriate for some types of data than for others. For this analysis, a measure of linguistic distance would likely be most appropriate. That, however, is a direction for future work; for simplicity the analysis relies instead on a few measures typically available in a standard package: Euclidean distances, Pearson correlations, and cosines.

Non-hierarchical clustering reveals several interesting patterns as the number of clusters arbitrarily imposed increases from two to eight:

- *With two clusters*, all of the English informants separate into one cluster and all of the American informants into the other.
- *With three clusters*, the informants from the west and parts of the southeast of England form a cluster, and the southern American informants form another. The remaining cluster is composed of informants from the East Midlands, East Anglia, Middlesex, and Massachusetts.
- *With four clusters*, all of the Americans regroup into a single cluster, while the English informants split into three clusters: an eastern group including East Anglia, part of Middlesex, and parts of the East Midlands; a more central group that includes the rest of the East Midlands and the southeast, and a western group.
- *With five clusters*, the Massachusetts informants and the American southerners form separate and distinct groups, and remain so in subsequent clustering – all further reconfigurations involve only the English informants. The eastern English group from the previous clustering expands to include members of the central group, which shrinks accordingly to include mainly only southeastern informants. The western group remains unchanged.
- *With six clusters*, the expanded eastern English cluster splits in two, yielding a mainly East Anglian cluster and another encompassing most of the East Midlands. The southeastern and western groups remain essentially unchanged.
- *With seven clusters*, the two Devonshire informants split out into a single group, leaving the rest of the clusters largely unchanged from the previous pattern with six clusters.
- *With eight clusters*, three informants from Lincolnshire and Rutland form a separate group distinct from a larger East Midlands group, into which the Middlesex informants move. The remaining East Anglian and southeastern informants form two distinct groups. The large western group breaks into northern and southern clusters, with the Devonshire informants, previously separate, joining the northern (or West Midlands) cluster.

Hierarchical analysis of the data set yields further insights. Typically, the choice of method has more effect on the clustering than the distance measure;

different clustering methods can yield quite different groupings, while different distance measures often yield rather similar ones. All approaches cluster the Massachusetts informants into a single cluster and the southern American informants into another. The southern American cluster often divides into two or three subclusters, and those subclusters tend to divide along regional lines: that is, one cluster will tend to be composed mainly of informants from West Virginia and southwestern Virginia, while the other(s) will tend to be composed mainly of informants from eastern Virginia and North Carolina. However, in every such case some westerners cluster with easterners and vice versa, with no obvious pattern involving gender, age, or type of speech.

Nine of the twelve hierarchical analyses – all except those using the single-linkage method – place the Massachusetts and southern American clusters in a larger cluster that includes most of the eastern English informants, while the western English informants form a separate broad cluster. Under most of those approaches, the Americans form a separate large subcluster and the eastern English informants form a separate subcluster, which is itself divided into a number of subclusters – usually three. In a few cases, one or another American group clusters with one or another of the eastern English subclusters. Using one method – average linkage within groups – and using any of the three distance measures, the southeastern English subcluster groups together with the southern Americans.

The eastern English informants tend to divide into three subclusters under most approaches: a group mainly including informants from the East Midlands and the area to the north of London, another mainly composed of East Anglians, and the last centered in the counties southeast of London, but tending to include a handful of informants north and west of London. The western English tend to divide into two groups, one composed of informants from the southwestern coastal counties and the other including most of the informants to the north and west of London. Although the southwestern coastal informants form a stable group, the other cluster is rather unstable: under almost every approach, at least some of the more northerly westerners cluster into the coastal group, but which ones do so varies by approach. The most westerly informants, in Devonshire, sometimes cluster into the coastal group, sometimes into the more northerly group, sometimes by themselves as an outlier cluster, and under one clustering method, along with the southeasterners and American southerners. Using the single linkage approach, the westerners form a single large cluster with no distinctive subclusters, except for the Devonshire informants, who form an outlier cluster distinct from all other English and American speakers.

Taken together, the clustering process thus yields a great deal of information about the similarity of informants among regions. Southern England has two broad groups of dialects, one eastern and one western. American speakers

are clearly distinct from most southern English speakers, but they appear to have more in common with eastern English speakers than western ones. The American southerners form a distinct group that, compared with the southern English speakers, is so uniform that speakers from the North Carolina coast can hardly be distinguished from those in southern West Virginia. East Anglians form a largely distinct group; speakers in the East Midlands form another, and speakers in the southeast yet another. Speakers from near London appear to have affinities with those of the southeast but also with those of the East Midlands. Southeastern speakers seem to be situated between east and west, linguistically speaking, but closer to the east. Western English speakers may be thought of as forming three rather indistinct groups, one in Devonshire, a more distinct one along the south coast, and a third making up the southwest Midlands – the Cotswolds and the Upper Thames and Severn Valleys.

Drawing on all twelve approaches and using majority (or, where necessary, plurality) rule, I assign the southern English informants to six regions, as shown in Table 3.1 and delineated in the map fo England at the bottom right of Figure 3.1: the East Midlands (EM), East Anglia (EA), the Southeast (SE), the Southwest (SW), Devonshire (DV), and the West Midlands (WM). The three American regions are eastern Massachusetts (MA), eastern Virginia and North Carolina (EV), and western Virginia and southern West Virginia (WV). The English regions broadly correspond to the dialect regions noted by Kurath and Lowman (1970) and are also largely congruent with the dialect regions delineated in Trudgill (1999) as well as with the dialect clusters that the author has found in unpublished cluster analyses of data from Orton and Dieth (1962).

The second rightmost column of Table B.2 in Appendix B provides a measure of how often informants cluster into their designated region under the approaches used here. While most informants in most English regions always cluster into the designated region, nearly every region has several informants that appear only loosely connected to it. For instance, the most northerly informant – designated 'Lincolnshire 1' – is clearly more like an East Midlands speaker than like those of any of the other regions in the analysis, but he is really a North Midlands speaker and thus tends to appear as an outlier under most approaches. (Under a few approaches he even clusters as an outlier in the Massachusetts group.) The regional classification of several informants near the borders of regions (especially near London, in Middlesex, Hertford, and Essex) is quite sensitive to the choices of approach and distance measure. That suggests that the regions are best thought of as rather loosely bounded: speakers appear to inhabit a linguistic continuum as much as they do a set of sharply distinct dialect regions. Indeed, informants at the borders of regions occasionally are classified into American groups, bringing to mind the hypothesis that American speakers themselves may best be thought of, like the border

informants, as having characteristics of several of the English regions.

Note also that London itself emerges as a center surrounded by rather different dialect regions. As shown in the last column of Table B.2, with the exception of Devonshire, every one of the regions assigned in the analysis has at least one informant within 40 miles of London. Even at such close distances, however, the English speakers tend to cluster rather regularly into separate clusters rather than into a metropolitan cluster, suggesting that the very extensive and protracted historical migrations from all over the British Isles to London had relatively little effect on the speech patterns of rural speakers in the surrounding region.[11]

3.5 Analyzing the Distribution of Variants within and among Regions

A few simple summary measures describing the distribution of variants among regions provide a remarkably clear picture of the extensive variation within and among regions and overlap across regions. Of the total of 288 different variants found in the sample, 91 percent were found somewhere in southern England; 20 percent are found only in southern England and are absent in the American regions. These two statistics alone imply that 22 percent of the variants found in southern England were not transplanted to (or were lost in) the American regions, suggesting a significant amount of leveling in the development of American regional speech forms.

Similarly, 80 percent of the variants were found in one or more of the American regions, while only nine percent were found only in America. Thus the overwhelming majority of American variants were clearly also native to southern England even in the twentieth century. The fact that about 12 percent of American variants were not found in southern England by Lowman may be taken to indicate some degree of innovation in America, but that conclusion should be tempered by the observation that many of the apparent innovations are known to have existed in southern England in earlier periods or were found somewhere in England by other twentieth-century fieldworkers such as those from the *Survey of English Dialects*. Moreover, the fact that fully half of the apparent innovations are shared across all three American regions further suggests that they could very well have been in the inventory of speech forms imported to America.

Table 3.1 shows the percentages of the total population of variants found in each region and the percentages shared between regions. For example, the

[11] Unfortunately, Lowman does not appear to have interviewed any Cockneys. An informant who had acquired working-class London speech in the mid- to late-nineteenth century would have been an invaluable addition to the survey.

Table 3.1: Percentage of Total Variants Shared Between Regions

	EM	EA	SE	SW	DV	WM	MA	EV	WV
EM	74.3	57.6	49.0	47.9	29.5	58.3	48.3	47.6	41.0
EA	57.6	62.8	43.1	42.4	27.4	50.7	42.0	41.0	35.1
SE	49.0	43.1	54.2	42.0	26.4	46.9	39.9	37.5	32.3
SW	47.9	42.4	42.0	58.0	29.2	52.1	37.5	37.5	33.7
DV	29.5	27.4	26.4	29.2	35.8	32.6	24.7	22.9	22.2
WM	58.3	50.7	46.9	52.1	32.6	71.5	44.4	44.8	40.6
MA	48.3	42.0	39.9	37.5	24.7	44.4	59.0	45.8	39.2
EV	47.6	41.0	37.5	37.5	22.9	44.8	45.8	63.5	50.0
WV	41.0	35.1	32.3	33.7	22.2	40.6	39.2	50.0	54.9

Note: EM = East Midlands; EA = East Anglia; SE = Southeast England; SW = Southwest England; DV = Devonshire; WM = West Midlands; MA = Massachusetts; EV = Eastern Virginia and North Carolina; WV = Southwestern Virginia and Southern West Virginia.

Table 3.2: Percentage of Regions' Variants Shared Between Regions

	EM	EA	SE	SW	DV	WM	MA	EV	WV
EM	100.0	91.7	90.4	82.6	82.5	81.6	81.8	74.9	74.7
EA	77.6	100.0	79.5	73.1	76.7	70.9	71.2	64.5	63.9
SE	65.9	68.5	100.0	72.5	73.8	65.5	67.6	59.0	58.9
SW	64.5	67.4	77.6	100.0	81.6	72.8	63.5	59.0	61.4
DV	39.7	43.6	48.7	50.3	100.0	45.6	41.8	36.1	40.5
WM	78.5	80.7	86.5	89.8	91.3	100.0	75.3	70.5	74.1
MA	65.0	66.9	73.7	64.7	68.9	62.1	100.0	72.1	71.5
EV	64.0	65.2	69.2	64.7	64.1	62.6	77.6	100.0	91.1
WV	55.1	55.8	59.6	58.1	62.1	56.8	66.5	78.7	100.0

first number in the first column of Table 3.1 shows that 74.3 percent of all variants were found in use in the East Midlands (EM), each by at least one informant; the third number informs us that 49.0 percent of those variants were also found in the Southeast (SE), and the seventh number (like the first number in the seventh column) reveals that 48.3 percent of them were found in Massachusetts (MA).

In contrast, Table 3.2 shows, by column, the percentages of a given region's total variants that it shares with each other region (that is, the numerator of every value in a column in Table 3.2 is the total number of variants found in the specified region). The seventh number of the first column of Table 3.2 thus shows that the variants shared between the East Midlands and Massachusetts constituted 65.0 percent of all the variants found in the East Midlands, while the seventh number of the first row reveals that the same variants made up 81.8 percent of all the variants found in Massachusetts.

The diagonal elements of Table 3.1 show that the East and West Mid-

lands have much larger shares of all variants – 74.3 percent and 71.5 percent, respectively – than any other regions, while the rest (excluding Devonshire, which has only two informants) have between 54 percent and 63 percent. It may be useful to note that regions with larger samples usually have more variants. There is a nearly log-linear relationship between the sample size and number of variants for the English regions, that is, a nearly linear relationship between the natural log of the number of informants in a region and the natural log of the number of variants. There is a similarly log-linear relationship among the American regions, but with a lower average number of variants per speaker. The lower diversity of variants among American speakers, an indication of more uniform speech patterns, compared with speakers in England, is consistent with – indeed, may well be a result of – population bottlenecks and founder effects associated with the settlement of the American colonies. That is, the relatively small number of colonists who settled any given locality may not have brought the full diversity of English variants with them, and the variants that survived into the second and third generations may have had a survival advantage over the variants used by subsequent newcomers.

The proportions in the last three columns of Table 3.1 show that American regions tend to share a larger number of variants with the East and West Midlands than with other regions. (That pattern, however, is not very robust: moving the speakers in Buckinghamshire into the East Midlands group, on which they border, dramatically reduces the percentage of variants found in the West Midlands and shared with American regions. With that minor change, American regions share more variants with eastern regions than with western ones.) Intriguingly, Massachusetts and Eastern Virginia share roughly the same proportion of their variants with nearly every English region, despite the fact that their mutually shared variants constitute only 77.6 percent of all variants in use in Massachusetts and only 72.1 percent of all those in Eastern Virginia. The comparison is somewhat complicated by the fact that there are more than 50 percent more informants in the Eastern Virginia group than in either of the other two American regions, in effect creating a larger environment for variants to coexist. Nevertheless, those numbers leave the impression that both American regions experienced a rather similar degree of influence from each English region and perhaps a similar degree of leveling, but with a different mix of variants resulting in each region. Both regions share more variants with each English region than does the Western Virginia region. The latter region has a noticeably lower population of variants than either of the other American regions, but shares 91.1 percent of its variants with Eastern Virginia, possibly indicating further leveling during the expansion process following initial colonization.

Another interesting observation – not shown in the tables – is that the southern American regions show a slightly greater affinity with those of the

western regions of England than does Massachusetts. That affinity can best be isolated by comparing the distributions of variants appearing in England exclusively in the east or in the west. Thirty-five variants appear in the eastern regions of England but not the western ones, while 25 appear in the west but not in the east. Of the purely eastern variants, 49 percent appear in Massachusetts and 69 percent in the American south. (Thirty-four percent appear both in Massachusetts and in the south, 14 percent only in Massachusetts, and 34 percent only in the south.) In contrast, only 20 percent of the purely western variants appear in Massachusetts, but 40 percent appear in the south. (Twelve percent appear in Massachusetts and in the south, eight percent only in Massachusetts, and 32 percent only in the south.)

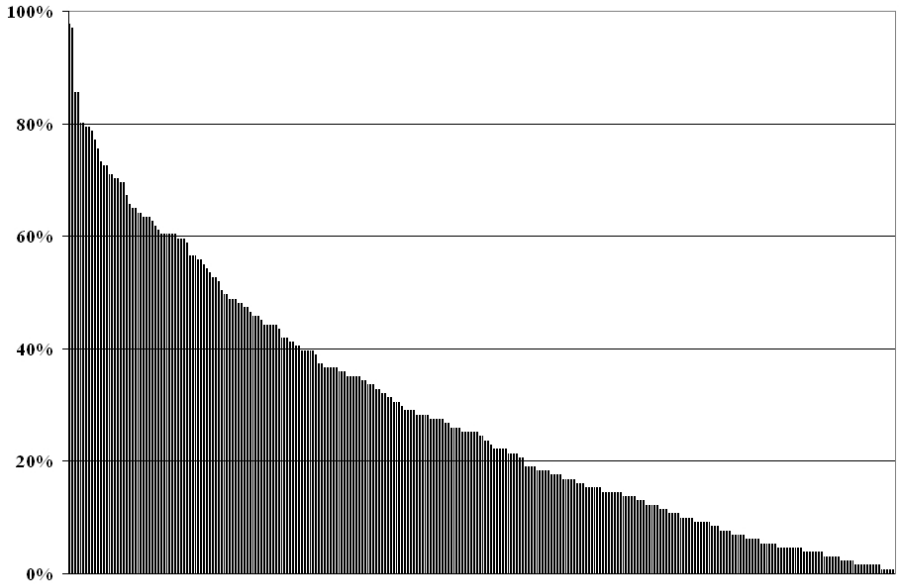
Some variants in the sample are very widespread, while others are rare. Over ten percent of the variants were recorded in all six English and three American regions. Nearly 17 percent were found in all six English regions; 40 percent in five of them, and 56 percent in four. Thirty-seven percent were found in all three American regions – further evidence suggestive of leveling across regions. Twenty-four percent were recorded in at least eight regions; 36 percent in seven regions; 51 percent in six; and 62 percent in five. Only about five percent of variants were recorded in only one region, another nine percent were found in only two. The distribution among informants mirrors that among regions: eight percent of the variants were used by over 75 percent of all informants, and one was used by nearly 98 percent. At the other end of the spectrum, 14 percent of the variants were recorded in use by less than five percent of all speakers, and 26 percent were used by less than ten percent.

Figure 3.2 illustrates the overall distribution of variants among informants in the sample, ranked from most widely to least widely in use. The pattern, resembling the A-curve or asymptotic hyperbolic distribution discussed by Kretzschmar and Tamasi (2002) and others, is a familiar one to dialectologists and is found for a wide range of linguistic phenomena. However, the distribution shown in the figure is much more linear or "flatter" than the pattern that is commonly found in such data. The reason for such apparent flattening is not obvious, but one plausible explanation is that the classification of variants by Kurath and McDavid obscures enough of the "actual" variation in the data to leave the impression that there are fewer very uncommon variants than is truly the case.

3.6 Measuring Degrees of Similarity among Informants: Shared Variants

The distribution of variants can be analyzed further by calculating not only the percentage of variants shared between regions, as in the previous section, but

Figure 3.2: Percentage of Informants Using Each Variant



also the percentage shared between individual informants. Table 3.3 shows the average proportion of shared variants between two randomly chosen informants in regions. The first number in the first column of Table 3.3, for instance, shows that on average, two informants in the East Midlands share 61.3 percent of their variants in common. The fourth number in the column shows that on average, an East Midlands informant shares only 35.6 percent of his variants with an informant picked at random from the Southwest – nearly the same percentage he shares with a random informant from the western Virginia region, as shown at the bottom of the column. (Since each informant may have more than one variant for a given phoneme in a given context, or may not have provided a response, informants may have different numbers of total variants. In that case the number of variants they share will be the same for both, but the percentage of their own variants that they share with each other may differ between the two informants. Thus the mean percentage of shared variants that an East Midlands informant shares with a Western Virginian is 35.4 percent, but the percentage that the Western Virginian shares with the East Midlands informant is 36.6 percent, as shown by the first element of the last column.)

Table 3.3: Mean Percentage of Shared Variants Between Speakers in Regions

	EM	EA	SE	SW	DV	WM	MA	EV	WV
EM	61.3	48.6	50.8	35.3	40.4	39.3	44.2	37.9	35.6
EA	48.8	64.5	41.9	36.1	35.6	36.9	40.4	38.8	37.6
SE	47.8	39.2	70.2	43.0	41.2	43.7	43.4	38.4	39.5
SW	35.6	35.7	45.0	67.3	53.4	53.9	32.0	35.3	39.7
DV	39.1	34.3	42.4	48.0	88.6	47.2	36.1	34.1	41.0
WM	39.3	36.7	46.5	55.2	48.7	62.7	32.0	33.0	36.0
MA	43.8	39.9	45.8	31.6	37.0	31.7	70.7	45.6	43.4
EV	37.4	38.2	40.4	35.3	34.9	32.7	45.5	71.3	66.0
WV	35.4	37.1	41.7	39.5	42.0	35.7	43.5	66.2	74.7

Note: EM = East Midlands; EA = East Anglia; SE = Southeast England; SW = Southwest England; DV = Devonshire; WM = West Midlands; MA = Massachusetts; EV = Eastern Virginia and North Carolina; WV = Southwestern Virginia and Southern West Virginia.

Table 3.4: Percentage of Shared Variants Between Typical Speakers in Regions

	Rt5	Sf25	Sr42	Hp59	Dv68	Gl81	MA 119.2	NC2B	V75A
Rutland 5	100.0	52.1	48.0	36.5	38.0	34.2	55.4	35.6	30.9
Suffolk 25	51.3	100.0	37.8	36.4	34.7	30.3	39.7	39.6	38.2
Surrey 42	43.7	34.9	100.0	42.4	35.4	45.0	52.0	38.6	32.3
Hamp. 59	36.3	36.7	46.3	100.0	43.9	61.9	34.5	42.8	42.2
Devon. 68	36.6	33.9	37.5	42.5	100.0	51.9	36.6	37.9	41.6
Glouc. 81	34.3	30.8	49.6	62.4	54.0	100.0	30.7	36.3	35.2
Mass. 119.2	56.1	40.7	57.8	35.0	38.5	31.0	100.0	52.1	38.3
No. Car. 2B	31.7	35.8	37.7	38.3	35.0	32.2	45.8	100.0	59.3
Virg. 75A	30.5	38.2	35.0	41.9	42.6	34.6	37.4	65.8	100.0

Table 3.4 shows the proportion of shared variants between the most "typical" informants in each region, defined as the informants having the highest average percentage of shared variants with all of the others in their respective regions. (The values of 100 percent in the diagonal elements indicate that a region's most typical informant shares all of his variants with himself.) As in Table 3.3, the value may vary between informants depending of which informant's number of variants is in the denominator.

The tables show that there is extensive variation within and among regions; again, informants appear to inhabit more of a linguistic continuum than a set of sharply delineated dialect regions. Informants in a given region typically share 60 percent to 75 percent of their variants – with the exception of Devonshire, whose two informants share nearly 89 percent of their variants – but the range within regions (not shown here) is 32 percent to 92 percent. The English regions' generally lower percentages indicate greater internal variation – and

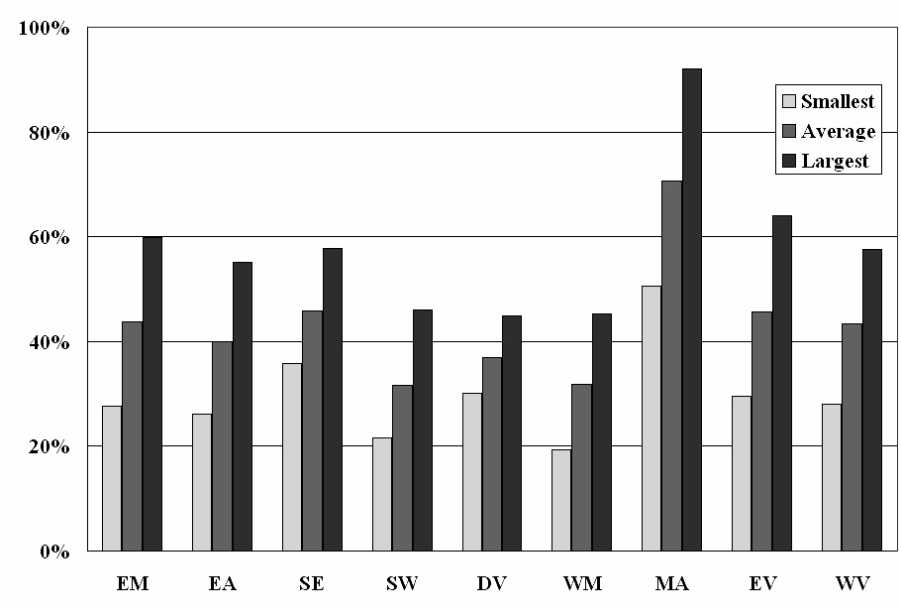
usually more variation amongst each other – than is the case for the American ones, which are relatively homogeneous internally and also relatively similar. A randomly chosen pair of English informants may share as many as 83 percent of their variants or as few as 22 percent; a similarly chosen pair of American informants may share any many as 92 percent or as few as 28 percent. The distinction between eastern and western English speech patterns shows up very clearly in the tables: informants in the East Midlands and East Anglia typically share less than 40 percent of their variants with informants from the Southwest, Devonshire, or the West Midlands. The affinity of the Southeastern region with both east and west is also apparent in the third column of each table. The three speakers in Middlesex reveal substantial variation (not shown in the tables) in the vicinity of London; one pair of them shares only about 57 percent of their variants, indicating greater diversity between them than is typical between informants within any English region.

The ranges of variation between the English and American regions are generally similar. English and American informants typically share 35 percent to 40 percent of their variants, but (not shown) may share as many as 60 percent or as few as 19 percent. Most importantly, the results indicate that the American speech forms fall squarely into the family of southern English speech varieties. That is, American informants typically share as many variants with the southern English informants as the English informants share with each other. For instance, the range of average shared variants between the East Midlands informants and American informants – 35 percent to 44 percent – is very similar to that between East Midlands informants and western English ones – 36 percent to 39 percent; indeed, it is even somewhat larger.

Informants from Massachusetts generally share more variants with eastern English (particularly East Midlands) informants and fewer variants with western informants than do American southerners. The most typical Massachusetts informant shares more variants with the most typical East Midlands and Southeastern informants than nearly any other pair of typical regional informants. In contrast, southern American informants – who are comparatively homogeneous as well as similar in their intra- and interregional variation – have more diffuse affinities in general than do the Massachusetts informants. On average, the typical informants from the southern American regions have greater similarity with their counterparts in the western English regions than does the typical Massachusetts informant. That distribution of shared variants strongly suggests that different populations of variants and leveling processes among North American settlers produced somewhat different populations of variants in different regions.

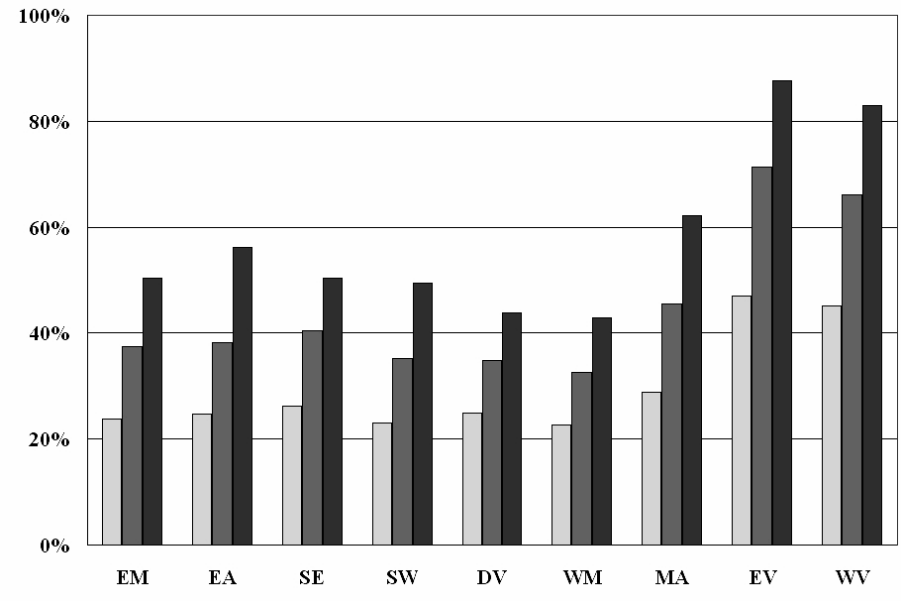
Figures 3.3 through 3.5 illustrate those observations by showing the percentages of variants shared between each of the American regions and all of the other regions. For each comparison, the lightest bar shows the smallest

Figure 3.3: Percentage of Shared Variants: Massachusetts



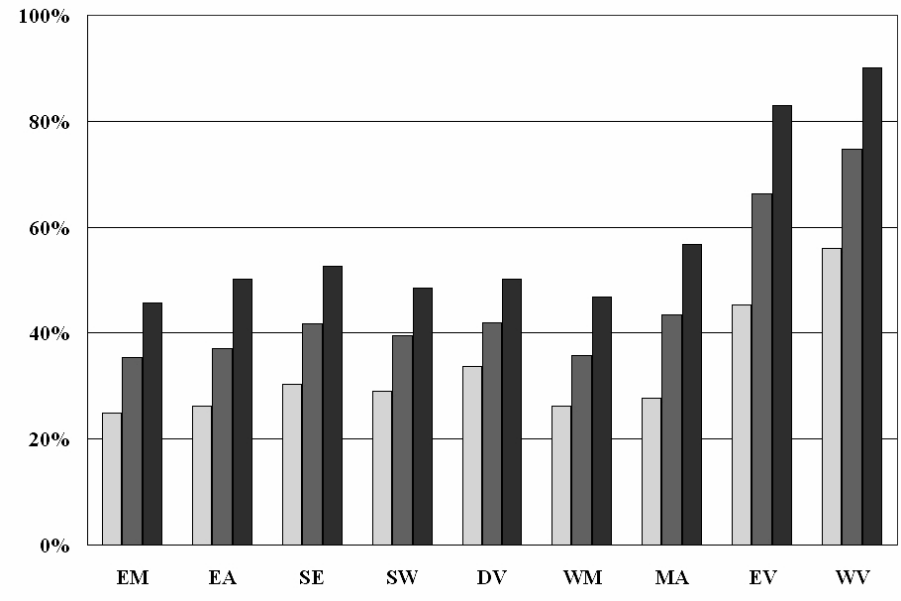
Note: EM = East Midlands; EA = East Anglia; SE = Southeast England; SW = Southwest England; DV = Devonshire; WM = West Midlands; MA = Massachusetts; EV = Eastern Virginia and North Carolina; WV = Southwestern Virginia and Southern West Virginia.

Figure 3.4: Percentage of Shared Variants: Eastern Virginia



Note: EM = East Midlands; EA = East Anglia; SE = Southeast England; SW = Southwest England; DV = Devonshire; WM = West Midlands; MA = Massachusetts; EV = Eastern Virginia and North Carolina; WV = Southwestern Virginia and Southern West Virginia.

Figure 3.5: Percentage of Shared Variants: Western Virginia



Note: EM = East Midlands; EA = East Anglia; SE = Southeast England; SW = Southwest England; DV = Devonshire; WM = West Midlands; MA = Massachusetts; EV = Eastern Virginia and North Carolina; WV = Southwestern Virginia and Southern West Virginia.

percentage shared between an informant in the American region and another in the specified region; the middle bar shows the average, and the darkest bar shows the largest.

3.7 Measuring Degrees of Similarity among Informants: Linguistic Distances

Entirely different – and, perhaps, more linguistically relevant – measures of similarity can be constructed by translating variants into vectors of numerical values representing degrees of height, backing, rounding, rhoticity, length, and so forth, and by measuring linguistic distance as a Euclidean distance between variants in an idealized geometric grid (e.g. [ɛ] and [e] are closer to each other than [i] and [a]).[12] To measure linguistic distance in the sample used in this study, each short vowel is represented as a vector of four numbers, each representing a feature of the vowel: one to three for the degree of backing, one to seven for height, one to two for rounding, and one to three for rhoticity. Long vowels and diphthongs are represented by a vector of eight values (half-lengthening is treated as full lengthening), and the distance between a short vowel and its lengthened twin (or a diphthong involving the short vowel) is taken to be 1. Distances between variants of consonants are generally given a value of 1. The vector characterization adopted here for each of the variants distinguished in the data sample is given in Table B.1. For the most part, those characterizations represent the mean value for each feature for the range of vowels included by Kurath and McDavid under each specific variant. For some variants, however – those that are designated "other" and may represent a hodgepodge of forms – the characterization is necessarily somewhat arbitrary.

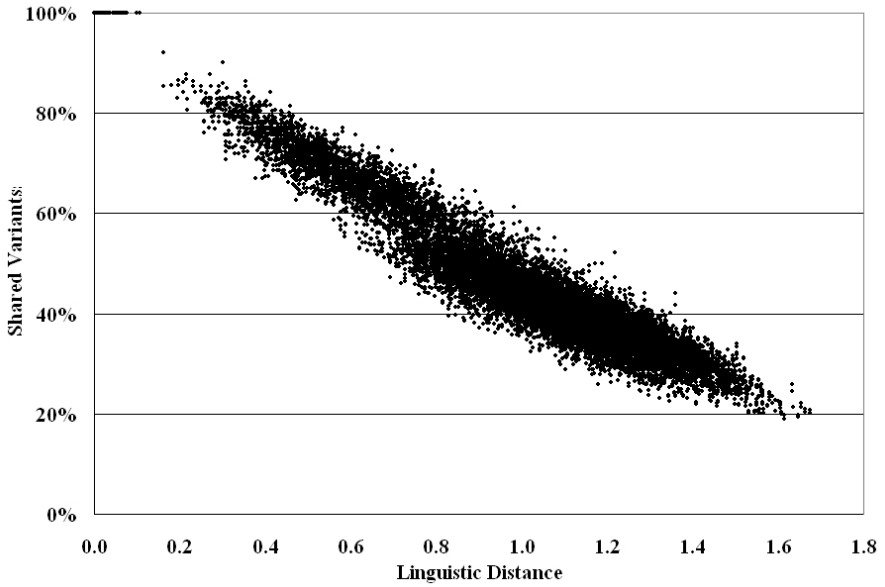
Linguistic distance between any two speakers is calculated as the average distance over all variants. Over the sample used in this study, linguistic distance takes values ranging from 0.0 to nearly 1.8. Linguistic distance thus provides an intuitive feel for the degree of difference between speakers' usages: a measure of 1.0 implies that on average, two speakers' phonemes typically vary as between [e] and [ɛ], or between [a] and [ɑ].[13]

The variance of linguistic distance – a measure of the degree of dispersal in the distances between two speakers' variants – also provides useful insights.

[12] See Heeringa (2004) – especially chapter three – for a highly detailed discussion of such approaches to measuring linguistic distance. The approach used in this study is most similar to that of Almeida and Braun, discussed on pp. 40-45.

[13] Note that in accordance with standard practice in the American *Linguistic Atlases*, [ɑ] denotes a low-central unrounded vowel in this chapter and the next, and the low-back unrounded vowel nornally denoted by [ɑ] in the International Phonetic Alphabet is denoted by [ɑ].

Figure 3.6: Linguistic Distance versus Shared Variants



The smaller the variance, the more the two speakers tend to have a large number of differences of similar size between their pronunciations; the larger that variance, the more two speakers tend to have a number of variants in common but a number of variants that are linguistically very different. Allowance is made for speakers having two variants. Hence a speaker may show a minor degree of distance from himself or herself, indicating a degree of variation in his or her speech pattern.

An important caveat to this approach is that any such linguistic measure involves a degree of arbitrariness in the conversion of perceptual qualities to numerical quantities. Moreover, as mentioned previously, the data used here suffers from having already been classified into groups of variants, with a consequent loss of variation and information. However, the disadvantages of arbitrariness in characterization and quantification appear to be outweighed by the advantages of being able to quantify, however imperfectly, a measure of perceptual or articulatory distance. Furthermore, such an approach allows one to take into account the largely continuous nature of linguistic phenomena.

That the linguistic distance measure provides additional information not captured in the shared variants measure can be seen in Figure 3.6, which

Table 3.5: Mean Linguistic Distance Between Speakers in Regions

	EM	EA	SE	SW	DV	WM	MA	EV	WV
EM	0.735	0.986	0.883	1.187	1.147	1.193	1.008	1.137	1.188
EA	0.986	0.679	1.056	1.186	1.184	1.231	1.117	1.213	1.217
SE	0.883	1.056	0.491	0.977	1.075	1.060	0.908	1.079	1.081
SW	1.181	1.187	0.990	0.652	0.911	0.932	1.260	1.221	1.165
DV	1.147	1.184	1.075	0.988	0.214	1.077	1.252	1.242	1.121
WM	1.193	1.231	1.060	0.909	1.077	0.758	1.333	1.307	1.269
MA	1.008	1.117	0.908	1.260	1.252	1.333	0.487	0.936	0.982
EV	1.137	1.213	1.079	1.219	1.242	1.307	0.936	0.511	0.627
WV	1.188	1.217	1.081	1.169	1.121	1.269	0.982	0.627	0.497

Note: EM = East Midlands; EA = East Anglia; SE = Southeast England; SW = Southwest England; DV = Devonshire; WM = West Midlands; MA = Massachusetts; EV = Eastern Virginia and North Carolina; WV = Southwestern Virginia and Southern West Virginia.

graphs each pair of informants' shared variants measures against the corresponding linguistic distance measures. Although the two measures are highly negatively correlated, with a coefficient of -0.96, the linguistic distance associated with any particular value of shared variants may vary by a factor of two. For instance, for a shared variants value of 50 percent, two speakers may have a linguistic distance of roughly 0.6, while another pair may have a linguistic distance of about 1.2. The intuition is that two speakers who have half of their variants in common (and zero average linguistic distance over those variants) may have variants elsewhere that are rather similar, or quite different, linguistically speaking. That variation may permit distinctions to be made among pairs (or groups) of speakers who share the similar percentage of their variants but who have major or minor differences among the variants that they do not share.

Table 3.5 shows the average linguistic distance between speakers in regions, while Table 3.6 shows the linguistic distance between the most typical speakers in each region, now defined as the speaker with the lowest sum of linguistic distances with all of the other speakers in that region. Table 3.7 shows the average standard deviation in the linguistic distance measures within and among regions.

The linguistic distance measures provide some further insights to those derived from the shared variants measures.[14] Massachusetts informants have

[14] Another set of distance measures are available from genetic research, which deals with problems that are in many ways analogous to those of historical linguistics and dialectology. Measures of genetic distance are typically based on the relative frequencies of different genetic variants or alleles of a given gene. Such measures can be applied to linguistic data, treating variants of a given phoneme as analogous with alleles of a given gene. One such measure, Nei's genetic distance (D), measures how closely related populations are under the assumption that change is always to a completely new variant, all genes have the same rate

Table 3.6: Linguistic Distance Between Typical Speakers in Regions

	Rt5	Sf26	Sr43	Ds65	Dv68	Ox84	MA 119.2	NC5A	V72B
Rutland 5	0.013	0.836	0.692	1.307	1.092	1.262	0.668	0.951	1.066
Suffolk 26	0.836	0.002	1.012	1.227	1.088	1.252	1.097	1.217	1.187
Surrey 43	0.692	1.012	0.033	0.975	1.102	0.946	0.757	1.079	0.970
Dorset 65	1.307	1.227	0.975	0.054	0.922	0.558	1.308	1.337	1.180
Devon. 68	1.092	1.088	1.102	0.922	0.047	1.010	1.199	1.281	1.140
Oxford 84	1.262	1.252	0.946	0.558	1.010	0.073	1.267	1.414	1.269
Mass. 119.2	0.668	1.097	0.757	1.308	1.199	1.267	0.000	0.859	0.989
No. Car. 5A	0.951	1.217	1.079	1.337	1.281	1.414	0.859	0.034	0.566
Virg. 72B	1.066	1.187	0.970	1.180	1.140	1.269	0.989	0.566	0.019

Table 3.7: Standard Deviation of Linguistic Distances Between Speakers in Regions

	EM	EA	SE	SW	DV	WM	MA	EV	WV
EM	0.275	0.147	0.123	0.117	0.099	0.195	0.124	0.107	0.109
EA	0.147	0.263	0.139	0.109	0.130	0.123	0.136	0.132	0.118
SE	0.123	0.139	0.220	0.131	0.078	0.199	0.112	0.098	0.113
SW	0.117	0.111	0.131	0.253	0.261	0.173	0.121	0.098	0.088
DV	0.099	0.130	0.078	0.098	0.192	0.105	0.070	0.095	0.080
WM	0.195	0.123	0.199	0.165	0.105	0.280	0.157	0.103	0.096
MA	0.124	0.136	0.112	0.125	0.070	0.157	0.161	0.128	0.162
EV	0.107	0.132	0.098	0.099	0.095	0.103	0.128	0.166	0.139

Note: EM = East Midlands; EA = East Anglia; SE = Southeast England; SW = Southwest England; DV = Devonshire; WM = West Midlands; MA = Massachusetts; EV = Eastern Virginia and North Carolina; WV = Southwestern Virginia and Southern West Virginia.

lower linguistic distances from eastern English informants than from western ones. Distances between southern American and English informants are relatively similar across regions, but compared with those of informants from Massachusetts, greater in the east and smaller in the west – with the exception of the West Midlands, whose informants have the greatest distance from and least linguistic similarity with American informants of any English region. Furthermore, the standard deviation measures are lower for the distances between southern Americans and western English informants than eastern informants: not only are the American southerners roughly as close to the English westerners as they are to the easterners, but their differences with westerners, by phoneme, are somewhat more uniform than their differences with easterners.

3.8 Using Principal Component Analysis to Uncover Linguistic Structure

Principal component analysis refers to a set of mathematical procedures for determining whether (and which) variables in a data set form coherent subsets.[15] Principal component analysis reduces the number of dimensions in the data set by finding groups of variables that tend to occur together and that are relatively independent from other groups. In that respect, they simplify the data by grouping variables in a way somewhat similar to that in which cluster analyses simplify it by grouping observations. Principal component analysis uncovers sets of variables that are strongly positively or negatively correlated, i.e. that tend to occur together or that always occur separately, and combines them into principal components that are essentially linear combinations of the correlated variables. (In this sense, variables that always occur separately are not independent – rather, independence implies that there is no pattern of co-occurrence at all.) Thus, a principal component typically has two "poles," one involving large positive values, or loadings, for a group of variables that

of change, and the populations remain constant in size over time. Exactly similar patterns of variants will yield a value of 0.00; two informants with 50 percent shared variants will yield a value of roughly 0.7; two informants with one shared variant will yield a value of about 4.4, and two entirely dissimilar informants yield an infinite value. Measuring D in the sample used in this study, yields values ranging from 0.00 to 1.70. An analysis of Nei's distances among informants yields essentially the same insights as obtained from the analysis of shared variants. The quality of the data apparently does not allow the greater sophistication of the technique to yield any more insight than can be gained from a more transparent measure. Moreover, the data characteristics that make Nei's distance most appropriate do not apply to the case of language change: Linguistic change need not involve shifts to entirely new variants or uniform rates of change; and the populations of speakers have certainly not remained constant over time.

[15] See Tabachnick and Fidell (2000), Chapter 13, for a useful overview of principal component and factor analysis.

tend to be found together, and another involving large negative loadings for a different group of variables that are also found together but never with the first group.

In conventional principal component analysis, each principal component is orthogonal to (uncorrelated with or independent of) every other. The first principal component "extracts" or accounts for the maximum possible variance from the data set that can be accounted for by a single linear combination of variables; the second principal component will extract the maximum possible amount of the remaining variance, and so on.[16] Observations – in this case, informants – can be assigned component scores on the basis of how strongly the variables of a principal component occur, thus providing a value for presence of the variables that receive high loadings in the principal component in that person's speech.

Applied to a data set of linguistic features, principal component analysis may isolate sets of linguistic features that tend to occur together and not with other features. Some of those groups may be readily explained in structural linguistic terms, and the principal component scores may reveal clusters of speakers in localities (or at least trends among regions) that anchor those linguistic structures in specific regions. Labov et al. (2006) provides an illuminating linguistic application in which the frequencies of first and second formants of various vowels in the speech of several hundred American speakers are subjected to principal component analysis. Labov's first principal component, accounting for about 22 percent of the total variation in his data set, assigns positive loadings to formant values indicative of the Northern Cities Shift and negative loadings to those indicative of the Southern Shifts. The second principal component, accounting for about 14 percent of the total variation, assigns positive loadings to formant values associated with the "split short [**a**]" system found in New York City and the Mid-Atlantic region, and negative loadings to format values indicative of no split. The two principal components thus help uncover from a highly variable data set a set of linguistic structural patterns that distinguish eastern from western speakers as well as northern and southern speakers.

Tables 3.8 through 3.11 show results for the first two principal components from a standard principal component analysis to the English-American data set. Only the first two principal components – representing about 24 percent of the total variance in the data set – yield any obvious linguistic significance,

[16] Technically, standard principal component analysis extracts maximum variance from a data set using orthogonal vectors projected through the data: each successive component minimizes the sum of squared deviations remaining after the previous one, subject to the constraint that the component be orthogonal to the previous one(s). Variants on standard principal component analysis that "rotate" the PCs allow for a trade-off between orthogonality of components and extraction of variance.

and the pattern does not appear to be robust to moderate shifts in approach. However, the linguistic significance of each principal component is very clear and, moreover, consistent with the preceding discussion. As shown in Table 3.8, the first principal component has its largest positive loadings for a set of linguistic features that tend to be found (though not exclusively or invariably) in the West Midlands and do not tend to be found either in eastern England or in America. They include:

- lack of merger between Middle English [**aː**] and [**ai**];
- lack of merger between Middle English [**ou**] and [**ɔ**];
- a centered onset in *nine*;
- a variant of [**e**] in *grease*;
- an ingliding diphthong in *Mary*, *bracelet*, and other contexts;
- a low, centered or slightly fronted [**a**] in most contexts in which Americans use [**æ**];
- a low, generally unrounded vowel in contexts in which Americans typically use a higher and more rounded one: *because*, *daughter*, *law*, *haunted*, *forty*, and *joint*; and
- loss of [**h**] in initial position.

The first principal component has contrastingly large negative loadings for a set of features that tend (but, again, are not exclusively or invariably) to be found in America and in the east of England, including:

- merger between Middle English [**aː**] and [**ai**], and between Middle English [**ou**] and [**ɔ**];
- consistent with the latter merger, a variant of [**ɛɪ** ∼ **eɪ**] in both *bracelet* and *day*;
- a raised [**æ**] not only as the typical expression of the low, fronted vowel but also in *married*, *parents*, *haunted*, and *chair*;
- more retracted, rounded vowels in *boiled*, *joint*, *daughter*, and *haunted* (though not in *forty*); and
- retention of [**h**] in initial position.

Table 3.8: Loadings for the First Principal Component (Values > |0.32|)

Positive Loadings		Negative Loadings	
DSSE Fig. 16: other than **æ** before **p**, **t**, **g**, **k**, **n**, **r**	0.807	Map 106: variant of **e** in *tomato*	-0.861
DSSE Fig. 32: **h** lost	0.782	Map 165: **tjuz** in *Tuesday*	-0.812
Map 106: **a** ∼ **ɑ** ∼ **ɑ** in *tomato*	0.741	Fig. 16: **æ** before **p**, **t**, **g**, **k**, **n**, **r**	-0.797
DSSE Fig. 17 and *PEAS* Map 14: other than **æ** before fricatives	0.741	*DSSE* Fig. 32: **h** retained	-0.782
Map 51: variant of **ɑ** in *married*	0.722	*DSSE* Fig. 17 and *PEAS* Map 14: **æ** before fricatives	-0.741
Map 50: variant of **ɛə** ∼ **eə** in *Mary*	0.694	Map 19: **eɨ** ∼ **ɛɨ** in *bracelet*	-0.739
Map 26: centered onsets in *nine*	0.672	Map 24: **ɒɔ** ∼ **ɔo** in *dog*	-0.716
DSSE Fig. 9: Middle English **ou** not merged with Middle English **ɔ:** into an upgliding diphthong	0.651	Maps 102-104: **æ** in *parents*	-0.702
Map 143: **ɑɨ** ∼ **ɑɨ** ∼ **ɐɨ** in *joint*	0.650	Map 75: **æ** in *hammer*	-0.674
Map 171: **s** in *greasy*	0.643	Map 51: **æ** in *married*	-0.663
Map 133: **ɑ** ∼ **a** in *because*	0.624	Maps 18-19: Merger of Middle English **a:** and **ai**	-0.655
Map 129: variant of **a** ∼ **ɑ** ∼ **ɑ** in *daughter*	0.616	*DSSE* Fig. 9: Middle English **ou** not merged with Middle English **ɔ:** into an upgliding diphthong	-0.651
Map 22: **ɑ:** ∼ **ɑ:** in *law*	0.615	Map 18: **eᵊ** ∼ **ɛᵊ** in *day*	-0.647
Map 19: **eᵊ** ∼ **ɛᵊ** in *bracelet*	0.612	Map 131: **æ** in *haunted*	-0.646
Map 169: **v** in *nephew*	0.600	Map 144: **ɔɨ** in *boiled*	-0.645
Map 125: **ʌ** ∼ **ɵ** in *won't*	0.583	Map 98: **i** in *neither* or *either*	-0.644
Map 98: **ai** in *neither* or *either*	0.579	Maps 114-15: **ʌ** in *soot*	-0.638
Map 144: variant of **oɨ** etc. in *boiled*	0.579	Map 171: **z** in *greasy*	-0.631
Map 43: **u** ∼ **ʊ** in *four*	0.565	Map 123: variant of **o** in *home*	-0.611
Map 75: **a** in *hammer*	0.564	Map 169: **f** in *nephew*	-0.577
Maps 114-15: **ʊ** in *soot*	0.562	Map 129: variant of **ɔ** ∼ **ɒ** in *daughter*	-0.550
Map 32: **a:** ∼ **a:ˀ** in *father*	0.560	Map 143: **ɔɨ** in *joint*	-0.543
Map 131: **a** ∼ **ɑ** ∼ **ɑ** in *haunted*	0.540	Map 40: **æ** in *chair*	-0.535
Map 42: **u** ∼ **ʊ** in *poor*	0.537	Map 164: **ju** in *new*	-0.531
Map 45: variant of **ɑ** ∼ **ɑ** in *forty*	0.521	Map 45: variant of **ɔ** ∼ **ɒ** in *forty*	-0.521
DSSE Fig. 4: **eə** ∼ **e:** in three words with Middle English **ea**	0.510	Maps 161-62: **laibɛri** for *library*	-0.519
Map 164: **ju** in *new*	0.506	*DSSE* Fig. 4: other than **eə** ∼ **e:** in three words with Middle English **ea**	-0.510
Maps 18-19: No merger of Middle English **a:** and **ai**	0.504		

Table 3.9: Component Scores for the First Principal Component

Location	Score	Location	Score	Location	Score
Hampshire 58	1.485	Suffolk 25	0.817	*North Carolina 2B*	-1.045
Goucestershire 81	1.440	Hertfordshire 37	0.734	*West Virginia 30B*	-1.052
Warwickshire 90	1.428	Huntingdonshire 15	0.706	*West Virginia 31A*	-1.054
Oxford 84	1.395	Sussex 50	0.702	*Virginia 70B*	-1.063
Warwickshire 88	1.393	Suffolk 23	0.700	*Virginia 71B*	-1.070
Gloucestershire 80	1.382	Essex 30	0.699	*West Virginia 29A*	-1.072
Sussex 49	1.373	Lincolnshire 2	0.653	*Virginia 75B*	-1.076
Wiltshire 61	1.364	Suffolk 26	0.608	*West Virginia 29B*	-1.077
Northamptonshire 12	1.333	Middlesex 34	0.599	*West Virginia 31B*	-1.090
Hampshire 57	1.327	Middlesex 33	0.579	*Virginia 72A*	-1.099
Oxford 86	1.311	Essex 29	0.549	*Virginia 37*	-1.110
Goucestershire 78	1.306	Middlesex 35	0.538	*North Carolina 10B*	-1.111
Oxford 83	1.306	Cambridgeshire 18	0.507	*Virginia 67A*	-1.129
Oxford 85	1.285	Suffolk 24	0.411	*Virginia 74A*	-1.134
Northamptonshire 10	1.263	Essex 31	0.395	*North Carolina 6*	-1.144
Surrey 44	1.248	**Massachusetts 116.1**	0.199	*Virginia 67B*	-1.148
Buckinghamshire 41	1.242	**Massachusetts 146.2**	0.184	*Virginia 73*	-1.148
Buckinghamshire 40	1.233	**Massachusetts 116.2**	0.168	*Virginia 75A*	-1.153
Sussex 48	1.215	**Massachusetts 119.2**	0.075	*Virginia 72B*	-1.157
Dorsetshire 65	1.160	**Massachusetts 120.1**	0.059	*North Carolina 3A*	-1.160
Bedfordshire 13	1.120	**Massachusetts 113.1**	0.056	*Virginia 71A*	-1.176
Hampshire 59	1.120	**Massachusetts 146.1**	0.017	*Virginia 40A*	-1.181
Dorsetshire 64	1.116	**Massachusetts 112.2**	0.014	*North Carolina 7B*	-1.182
Devonshire 69	1.074	**Massachusetts 122.1**	0.004	*North Carolina 4A*	-1.189
Kent 47	1.031	**Massachusetts 110.1**	-0.028	*North Carolina 5B*	-1.193
Worcestershire 92	1.030	**Massachusetts 118.1**	-0.046	*Virginia 36B*	-1.195
Lincolnshire 1	1.030	**Massachusetts 119.1**	-0.078	*Virginia 39*	-1.197
Northamptonshire 8	0.991	**Massachusetts 117.1**	-0.086	*North Carolina 9A*	-1.200
Lincolnshire 3	0.988	**Massachusetts 122.2**	-0.087	*Virginia 38*	-1.210
Rutland 5	0.977	**Massachusetts 120.2**	-0.112	*North Carolina 10A*	-1.219
Kent 46	0.939	**Massachusetts 124.2**	-0.134	*Virginia 35B*	-1.231
Surrey 42	0.935	**Massachusetts 110.2**	-0.168	*Virginia 35A*	-1.234
Cambridgeshire 16	0.922	**Massachusetts 123.1**	-0.181	*North Carolina 3B*	-1.236
Somerset 74	0.921	**Massachusetts 124.1**	-0.223	*North Carolina 7A*	-1.242
Norfolk 22	0.920	**Massachusetts 123.2**	-0.239	*North Carolina 1*	-1.243
Surrey 43	0.917	**Massachusetts 124.3**	-0.365	*North Carolina 5A*	-1.259
Norfolk 20	0.904	**Massachusetts 112.1**	-0.432	*North Carolina 4B*	-1.270
Hertfordshire 38	0.889	*Virginia 40B*	-0.732	*North Carolina 12B*	-1.275
Devonshire 68	0.866	*West Virginia 30A*	-0.952	*North Carolina 12A*	-1.279
Somerset 75	0.864	*Virginia 36A*	-0.952	*North Carolina 11A*	-1.303
Kent 45	0.860	*Virginia 70A*	-0.963	*North Carolina 8A*	-1.309
Norfolk 21	0.852	*North Carolina 2A*	-0.974	*North Carolina 8B*	-1.329
Warwickshire 89	0.840	*Virginia 74B*	-1.023	*North Carolina 9B*	-1.358
Leicestershire 7	0.837	*North Carolina 11B*	-1.029		

The component scores for the first principal component, indicating the strength of the features in informants' speech, are shown in Table 3.9. Informants from the West Midlands and Southwest of England have the highest positive scores, with nearly all the lowest positive scores in England all involving locales somewhat to the northeast of London. The principal component yields near-zero scores in Massachusetts (bolded), and negative scores in the American south (italicized) – with eastern North Carolina locales generally earning the most negative scores. The first principal component, in sum, appears to distinguish a set of largely western English and often conservative features from largely eastern and typically innovative features. It also indicates that those eastern features are much more common in America than are the western ones; that is, all American dialects tend to be composed largely of variants that are found mainly in southeastern England.

In contrast to the first, the second principal component, shown in Table 3.10, has large positive loadings for a group of variants that tend to be found in Massachusetts and in the southeast of England:

- non-rhotic variants in *barn*, *door*, *thirty*, and *father*;
- an absence of palatalization in *new*, *ear*, *here*, *chair*, *Tuesday*, and *care*;
- in the short high backed vowels, [**u**] in *Cooper*, *coop*, and *hoop*, but [**ʊ**] in *broom*; and
- a fairly uniform, rounded and often relatively high vowel in *daughter*, *law*, *oxen*, *water*, *wash*, *forty*, *nothing*, *tomorrow*, and *dog*;

The second principal component has large negative loadings for variants that tend to be found in the west of England and, fairly commonly, in the southern American regions:

- rhotic variants in *barn*, *door*, and *thirty* – and in *walnut*;
- palatalization in *new*, *ear*, *here*, *beard*, and *care* – and even [**č**] in *Tuesday*;
- in the short high backed vowels, [**ʊ**] in *Cooper*, *coop*, and *hoop*, but [**u**] in *broom*; and
- in place of a uniform, rounded high vowel, two vowels; a low, back, unrounded vowel in *daughter*, *law*, *oxen*, *water*, *wash*, *forty*, and *tomorrow* (as isolated in the first principal component as well); but a higher unrounded [**ʌ** ∼ **ʊ**] in *nothing* and *dog*.

Thus the second principal component, like the first, shows a distinct east-west regional English division, but its distribution among Americans is much

Table 3.10: Loadings for the Second Principal Component (Values > |0.32|))

Positive Scores		Negative Scores	
Map 109: **u** in *Cooper*	0.723	Map 53: **ɑ** ∼ **ɑ** in *tomorrow*	-0.720
Map 156: **ə̣** in *door*	0.686	Map 109: **ʊ** in *Cooper*	-0.717
Map 22: **ɒː** ∼ **ɒːᵊ** in *law*	0.682	Map 34: **jɜ** in *ear*	-0.659
Map 15: **ɒ** ∼ **ɔ** in *oxen*	0.679	Map 15: **a** ∼ **ɑ** in *oxen*	-0.654
Map 164: **u** in *new*	0.641	Map 25: variant of **ɚ** in *thirty*	-0.644
Map 25: **ɜɪ** ∼ **ɞɪ** ∼ **ɜ̣ɪ** in *thirty*	0.638	Map 164: **ju** in *new*	-0.618
Map 31: **r**-less **aː** ∼ **ɑː˂** in *barn*	0.635	Map 108: **ʊ** in *coop*	-0.611
Map 151: **ə** in *father*	0.624	Map 45: variant of **ɑ** ∼ **ɑ**	
Map 29: **aʊ** ∼ **ɑʊ** in *out*	0.595	in *forty*	-0.586
Map 34: **i** ∼ **ɪ** in *ear*	0.587	Map 35: **jɜ** in *here*	-0.570
Map 45: variant of **ɔ** ∼ **ɒ**		Map 151: **ɚ** in *father*	-0.569
in *forty*	0.586	Map 178: **ɔrnʊt** ∼ **ɒʊnɪt**	
Map 134: **ɔ** ∼ **ɒ** in *water*	0.560	in *walnut*	-0.565
Map 108: **u** in *coop*	0.549	Map 156: **ɚ** ∼ **r** in *door*	-0.558
Map 88: **ɒ** in *nothing*	0.543	Map 36: **jɜ** in *beard*	-0.513
Map 53: **ɒ** in *tomorrow*	0.535	Map 22: **ɑː** ∼ **ɑː** in *law*	-0.504
Map 35: **i** ∼ **ɪ** in *here*	0.523	Map 39: **jɜ** in *care*	-0.494
Map 135: **ɔ** ∼ **ɒ** in *wash*	0.519	Map 153: **i** in *borrow*	-0.478
Map 129: variant of **ɔ** ∼ **ɒ**		Map 76: **a** ∼ **ɑ** in *radish*	-0.474
in *daughter*	0.498	Map 134: **a** ∼ **ɑ** ∼ **ɑ** in *water*	-0.470
Map 40: **ɛ** in *chair*	0.494	Map 107: **u** in *broom*	-0.469
Map 153: other than **i**		Map 129: variant of	
in *borrow*	0.478	**a** ∼ **ɑ** ∼ **ɑ** in *daughter*	-0.459
Map 17: **ʊu** ∼ **uː** ∼ **u** in *two*	0.475	Map 166: **ist** for *yeast*	-0.457
Map 107: **ʊ** in *broom*	0.469	Map 24: **ʌ** ∼ **ʊ** in *dog*	-0.453
Map 165: **tuz** in *Tuesday*	0.466	Map 133: **ɑ** ∼ **a** in *because*	-0.452
Map 39: **ɛ** in *care*	0.455	Map 40: **i** ∼ **ɪ** in *chair*	-0.448
Map 55: variant of **ɜr**		Map 135: **ɑ** ∼ **a** in *wash*	-0.445
in *furrow*	0.454	*DSSE* Fig. 30: voiced	
Map 24: **ɒː** ∼ **ɒːᵊ** in *dog*	0.452	fricative for **f**	-0.439
DSSE Fig. 30: unvoiced		Map 88: **ʌ** in *nothing*	-0.432
fricative for **f**	0.439	Map 165: **čuz** in *Tuesday*	-0.431
Map 123: **ɵ** in *home*	0.426	Map 148: **ə** in *careless*, etc.	-0.410
Map 148: other than **ə**		Map 110: **ʊ** in *hoop*	-0.405
in *careless*, etc.	0.410		
Map 110: **u** in *hoop*	0.405		
Map 166: **jist** for *yeast*	0.404		
Map 125: **ʊ** in *won't*	0.402		

different. As shown in Table 3.11, Massachusetts informants (bolded) have the highest positive component scores, followed by English informants in the East Midlands and particularly in the vicinity of London. Southwestern English informants take the largest negative scores, with southern American speakers (italicized) nearly all taking negative scores as well – sometimes larger scores than for southwestern English locales. The implications for American speech patterns are clear: the second principal component identifies a set of variants found primarily in both the English southwest and in the American south, and distinguishes them from a set of variants found in both the English southeast and in New England.

Principal component analysis of the data thus reveals two sets of oppositions involving fairly clear linguistic structural interpretations and distinct regional distributions. As represented by Lowman's informants, southern English speech has a fairly strong demarcation between east and west. American speech forms appear to draw from all over the region (and possibly from others as well). However, American forms tend to be similar to eastern English ones, on the whole, but northern American forms tend to be much more so, while southern American speech reveals significant western English affinities.[17]

The component scores for both principal components are illustrated in Figure 3.7, the first on the vertical axis, the second on the horizontal. Note that the English informants spread across the figure in a pattern roughly analogous to their geographic distribution, while the American speakers form two distinct clusters, one in a distinctly eastern position, the other, considered along the horizontal axis, positioned midway between east and west.

3.9 Using Multiple Regression to Assess the Importance of Geographic Distance

Multiple regression analysis, the workhorse of statistical analysis, refers to a set of statistical techniques that allow one to assess the relationship between a variable of interest – a dependent variable – and a number of other independent variables, allowing for interaction among the latter.[18] For example,

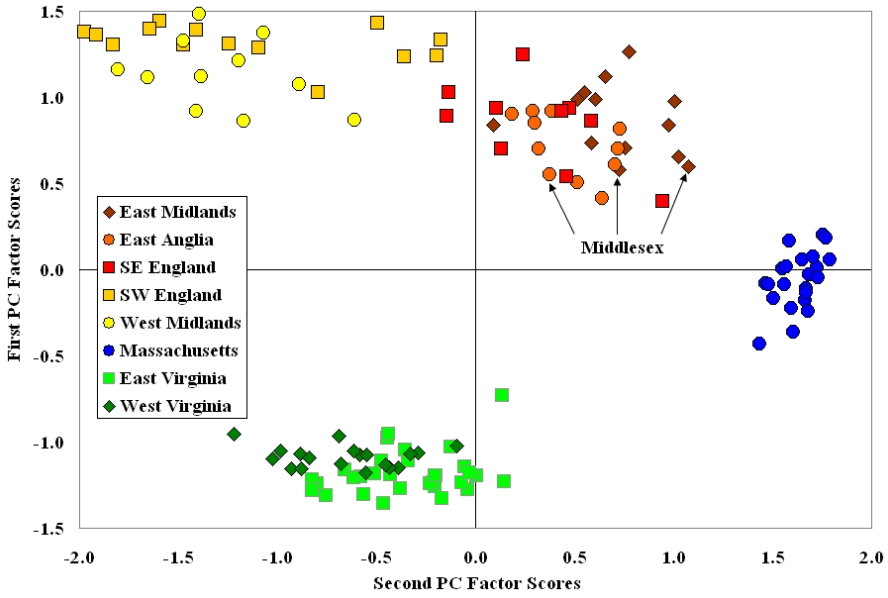
[17] The shortening of different words of the *coop/hoop/broom* family appears particularly diagnostic of English-American relations. Anderson (1987), on Map 18 on p. 36, documents the widespread tendency to shorten similar words throughout southern England, particularly in a belt from East Anglia to the West Midlands. However, different words tend to be shortened in different regions of southern England, and so it seems particularly interesting that the pattern of shortening varies across New England and the American South much as it varies from the southeast of England to the southwest.

[18] See Tabachnick and Fidell (2000), Chapter 5, for a useful overview of multiple regression analysis.

Table 3.11: Component Scores for the Second Principal Component

Location	Score	Location	Score	Location	Score
Massachusetts 120.1	1.789	Norfolk 22	0.384	*North Carolina 7B*	-0.510
Massachusetts 146.2	1.771	Essex 29	0.373	*West Virginia 29B*	-0.551
Massachusetts 116.1	1.756	Suffolk 23	0.322	*Virginia 71A*	-0.553
Massachusetts 118.1	1.729	Norfolk 21	0.299	*North Carolina 11A*	-0.564
Massachusetts 112.2	1.727	Cambridgeshire 16	0.288	*North Carolina 9A*	-0.579
Massachusetts 119.2	1.705	Surrey 44	0.238	*Virginia 75B*	-0.585
Massachusetts 110.1	1.684	Norfolk 20	0.184	Devonshire 68	-0.608
Massachusetts 123.2	1.680	*Virginia 35B*	0.145	*West Virginia 30B*	-0.615
Massachusetts 124.2	1.671	*Virginia 40B*	0.136	*Virginia 38*	-0.617
Massachusetts 120.2	1.669	Sussex 50	0.129	*North Carolina 3A*	-0.659
Massachusetts 123.1	1.665	Kent 46	0.104	*Virginia 67A*	-0.682
Massachusetts 113.1	1.648	Warwickshire 89	0.091	*Virginia 70A*	-0.690
Massachusetts 124.3	1.604	*Virginia 36B*	0.006	*North Carolina 8A*	-0.754
Massachusetts 124.1	1.597	*Virginia 40A*	-0.028	Worcestershire 92	-0.797
Massachusetts 116.2	1.586	*North Carolina* 12B	-0.042	*North Carolina 1*	-0.799
Massachusetts 146.1	1.568	*North Carolina* 6	-0.056	*North Carolina 12A*	-0.823
Massachusetts 117.1	1.558	*Virginia 35A*	-0.069	*North Carolina 10A*	-0.824
Massachusetts 122.1	1.552	*Virginia 74B*	-0.093	*West Virginia 31B*	-0.838
Massachusetts 110.2	1.506	*North Carolina 11B*	-0.125	*Virginia 75A*	-0.882
Massachusetts 122.2	1.478	Kent 47	-0.134	*West Virginia 29A*	-0.886
Massachusetts 119.1	1.467	Hertfordshire 38	-0.143	Devonshire 69	-0.891
Massachusetts 112.1	1.437	*North Carolina* 8B	-0.172	*Virginia 72B*	-0.929
Middlesex 34	1.076	Northamptonshire 12	-0.177	*West Virginia 31A*	-0.985
Lincolnshire 2	1.026	Buckinghamshire 41	-0.196	*Virginia 72A*	-1.027
Rutland 5	1.004	*North Carolina 5B*	-0.201	Sussex 49	-1.071
Leicestershire 7	0.973	*North Carolina 5A*	-0.205	Oxford 85	-1.094
Essex 31	0.944	*Virginia 39*	-0.207	Somerset 75	-1.172
Northamptonshire 10	0.775	*North Carolina 3B*	-0.213	Sussex 48	-1.197
Huntingdonshire 15	0.753	*North Carolina 7A*	-0.231	*West Virginia 30A*	-1.221
Suffolk 25	0.730	*Virginia 70B*	-0.291	Oxford 86	-1.246
Middlesex 33	0.723	*Virginia 71B*	-0.328	Hampshire 59	-1.384
Essex 30	0.722	*North Carolina 10B*	-0.339	Hampshire 58	-1.397
Suffolk 26	0.704	*North Carolina 2B*	-0.355	Warwickshire 88	-1.408
Bedfordshire 13	0.656	Buckinghamshire 40	-0.362	Somerset 74	-1.411
Suffolk 24	0.642	*North Carolina 4B*	-0.380	Oxford 83	-1.476
Northamptonshire 8	0.603	*Virginia 67B*	-0.390	Hampshire 57	-1.477
Kent 45	0.586	*North Carolina 4A*	-0.430	Goucestershire 81	-1.595
Hertfordshire 37	0.585	*Virginia 73*	-0.437	Oxford 84	-1.646
Lincolnshire 1	0.550	*Virginia 36A*	-0.441	Dorsetshire 64	-1.655
Lincolnshire 3	0.517	*North Carolina 2A*	-0.443	Dorsetshire 65	-1.807
Cambridgeshire 18	0.514	*Virginia 74A*	-0.457	Goucestershire 78	-1.829
Surrey 42	0.477	*North Carolina 9B*	-0.465	Wiltshire 61	-1.915
Middlesex 35	0.462	*Virginia 37*	-0.476	Goucestershire 80	-1.973
Surrey 43	0.435	Warwickshire 90	-0.495		

Figure 3.7: Principal Components for English and American Dialect Features



a researcher may use multiple regression to analyze how individuals' weight is simultaneously influenced by their height, waistline, and age.

Regression analysis can be used to test for a relationship between American informants' degree of similarity with English informants and the latters' geographic location. For instance, it may reveal a tendency for American informants to have more variants in common with English informants from nearest the metropolitan area and thus may lend support to the proposition that one cause of the relative uniformity of American speech – as well as its relative similarity to southeastern varieties of English – is the fact that many immigrants came from near London. According to Bailyn (1986), London was absorbing more or less all of the natural increase in population in Britain during at least part of the period of American colonization, and many migrants to America did so after first coming to London. (It is important to note, however, that they migrated to London, not to the rural areas in the vicinity of London. As mentioned previously, English informants living only a few score miles from London tend to cluster rather regularly into separate, distinct clusters rather than into a London-centered one, suggesting that population movements to London had relatively little effect on the speech patterns of rural speakers in

the surrounding region.)

Table 3.12 shows the results of a series of twelve regressions that apply such a test. In each regression, the values for one of the measures of similarity between all the informants from one of the American regions and all English informants is regressed against the distance in miles of the English informants' localities from London. The parameter labeled "Distance" provides an estimate of how much an increase in the English informants' distance from London changes an informant's similarity with American informants from the region.

For example, in the top left-hand regression labeled "Distance Only," the constant term indicates that according to the correlations in the data set, a hypothetical informant living in London would share about 42.6 percent of his or her variants with a typical Massachusetts informant. The parameter for the distance variable indicates that 100 miles of distance from London reduces an English speaker's proportion of shared variants with a typical Massachusetts speaker by about six percentage points. (The value of the parameter, -0.0006, times 100, yields -0.06 or minus six percentage points.) The low value of the "significance" measure paradoxically indicates that the value of the parameter is relatively well-constrained by the estimate. The adjusted R-square indicates that the regression accounts about for only about seven percent of the variance in the percentage of variants shared between Massachusetts and southern English informants. The standard error indicates that using this equation, the typical estimate of a Massachusetts informant's shared variants with an English informant will be off by nearly eight percentage points.

Another regression directly below the first, labeled "Regional Variables," includes regional "dummy" variables that take a value of 1 if the English informant is in a given region and 0 otherwise. That regression allows us to gauge the importance of distance from London while allowing for the fact that American informants may have different degrees of similarity with English informants from different regions. When dummy variables are used for a set of groups in a regression, one group is excluded, and the constant that is estimated in the regression is interpreted as the dummy for that group. In this case, the Southeast is excluded, and the constant term in the equation provides an estimate of the average degree of similarity between informants from the American region and a hypothetical Southeastern speaker living on the southern edge of London.

Table 3.12: Regression Results: Linguistic Similarity between American and English Speakers as a Function of Region and Distance From London

Massachusetts: Shared Variants			**Massachusetts: Linguistic Distance**		
(Distance Only)	Adjusted R-Sq.	Standard Error	**(Distance Only)**	Adjusted R-Sq.	Standard Error
	0.0700	0.0766		0.1550	0.1866
	Parameter	Significance		Parameter	Significance
Constant	0.4260	0.0050	Constant	0.9780	0.0110
Distance	−0.0006	0.0000	Distance	0.0024	0.0000
(Regional variables)	Adjusted R-Sq.	Standard Error	**(Regional variables)**	Adjusted R-Sq.	Standard Error
	0.5300	0.0544		0.5850	0.1308
	Parameter	Significance		Parameter	Significance
Constant	0.4590	0.0000	Constant	0.8840	0.0000
Distance	−0.0002	0.0000	Distance	0.0013	0.0000
East Midlands	−0.0067	0.1970	East Midlands	0.0425	0.0010
East Anglia	−0.0439	0.0000	East Anglia	0.1410	0.0000
Southwest	−0.1250	0.0000	Southwest	0.2870	0.0000
Devonshire	−0.0520	0.0000	Devonshire	0.1580	0.0000
West Midlands	−0.0069	0.1360	West Midlands	0.0838	0.0000
Eastern Virginia: Shared Variants			**Eastern Virginia: Linguistic Distance**		
(Distance Only)	Adjusted R-Sq.	Standard Error	**(Distance Only)**	Adjusted R-Sq.	Standard Error
	0.0370	0.0521		0.1000	0.1264
	Parameter	Significance		Parameter	Significance
Constant	0.3850	0.0030	Constant	1.1150	0.0070
Distance	−0.0003	0.0000	Distance	0.0013	0.0000
(Regional variables)	Adjusted R-Sq.	Standard Error	**(Regional variables)**	Adjusted R-Sq.	Standard Error
	0.2260	0.0467		0.3520	0.1073
	Parameter	Significance		Parameter	Significance
Constant	0.4030	0.0000	Constant	1.0620	0.0000
Distance	−0.0001	0.0050	Distance	0.0009	0.0000
East Midlands	−0.0204	0.0000	East Midlands	0.0224	0.0090
East Anglia	−0.0114	0.0050	East Anglia	0.0912	0.0000
Southwest	−0.0381	0.0000	Southwest	0.0912	0.0000
Devonshire	−0.0333	0.0000	Devonshire	0.0426	0.0330
West Midlands	−0.0329	0.0000	West Midlands	0.1070	0.0000

Continued on Next Page

Table 3.12 : Linguistic Similarity Regressed on Region and Distance (Continued)

Western Virginia: Shared Variants			**Western Virginia: Linguistic Distance**		
(Distance Only)	Adjusted R-Sq.	Standard Error	**(Distance Only)**	Adjusted R-Sq.	Standard Error
	−0.0010	0.0491		0.0250	0.1192
	Parameter	Significance		Parameter	Significance
Constant	0.3750	0.0030	Constant	1.1530	0.0080
Distance	0.0000	0.0000	Distance	0.0006	0.0000
(Regional variables)	Adjusted R-Sq.	Standard Error	**(Regional variables)**	Adjusted R-Sq.	Standard Error
	0.2130	0.0436		0.2290	0.1060
	Parameter	Significance		Parameter	Significance
Constant	0.4090	0.0000	Constant	1.0790	0.0000
Distance	0.0001	0.2720	Distance	0.0005	0.0000
East Midlands	−0.0592	0.0000	East Midlands	0.0790	0.0000
East Anglia	−0.0418	0.0000	East Anglia	0.1040	0.0000
Southwest	−0.0174	0.0000	Southwest	0.0525	0.0000
Devonshire	0.0017	0.8700	Devonshire	−0.0371	0.1410
West Midlands	−0.0395	0.0000	West Midlands	0.1090	0.0000

The second regression indicates that using the shared variants measure, all else being equal, the typical Massachusetts informant shares 45.9 percent of his or her variants with that hypothetical Southeastern English speaker. The parameter for the distance variable takes a much smaller value than in the previous regression, and now indicates that 100 miles of distance from London reduces an English speaker's proportion of shared variants with a typical Massachusetts speaker by about two percentage points, all else being equal, strongly suggesting that much of the variation accounted for by the distance parameter alone in the previous regression may be more appropriately accounted for by regional affiliation. Compared with informants from the English Southeast, informants from East Anglia, the Southwest and Devonshire will typically share significantly lower percentages of variants with Massachusetts informants than the English informants' distance from London alone would dictate – about 4.4, 12.5, and 5.2 percentage points lower, respectively. The significance values of practically zero indicate that those parameter estimates are well-constrained. In contrast, the parameters for the East Midlands and West Midlands dummy variables – which indicate that informants from those regions will typically share about two-thirds of a percentage point fewer variants with an informant from Massachusetts than a Southeastern informant living equidistant from London – are not significant, indicating that it is dif-

ficult to distinguish between the importance of distance for the Southeastern informants and those from the Midlands regions. In effect, distances to the west matter a great deal more than distances to the north and east, and the inclusion of regional variables brings that difference out of the data. The adjusted R-square indicates that distance and regional dummy variables account about for more than half of the variance in the percentage of variants shared between Massachusetts and southern English informants, with regional variations rather than distance from London accounting for most of that difference.

The shared variants regressions for southern American informants generally follow the same pattern. Regional location appears to be more important than distance from London in determining whether English informants have greater affinity with American ones. The regression for Eastern Virginians also yields a small but significant negative distance parameter estimate, again suggesting that American speakers do indeed have slightly greater similarity with rural speakers from near London. The parameter is much smaller, however, if regional variables are included in the regression, again suggesting that much of the variation accounted for by the distance parameter alone in the previous regression may be more appropriately accounted for by regional affiliation. The regression for Western Virginians, however, yields a zero distance parameter without regional dummies and a positive but insignificant one without it, indicating that distance from London has no independent affect on the similarity between English informants and American informants from that region. The negative parameters for all but one of the regional variables for the regressions that include such variables indicate that southern Americans typically share significantly fewer variants with informants in other regions than they do with informants from the Southeast (except, insignificantly, for Western Virginians compared to informants from Devonshire). Note that values of the Eastern and Western Virginians' regional parameters are rather similar to each other, compared with those of the New Englanders, and that the American southerners' regional parameters are more negative for the Midlands regions and less negative for the other regions than is the case for the New Englanders. Note also that the regressions for the American southerners have lower adjusted R-squares, indicating that regional affiliation and distance from London together account for less of the variation in their affinities with English informants, but that the regressions have lower standard errors than the ones for New Englanders, due to the relatively uniform character of their speech patterns. The results are entirely consistent with earlier findings: the American southerners show greater affinity with western speakers and less affinity with eastern speakers than do Massachusetts speakers, but their affinities are more diffuse altogether even though their speech is remarkably uniform.

The regressions on the linguistic distance measure yield essentially the same results, although the signs of the parameters are reversed because large values

for these measures indicate lesser rather than greater similarity. The addition of regional parameters affects the distance parameter estimate in a similar way; the regional parameters and their significances vary across English regions in a similar way; and the adjusted R-squares vary across American regions in a similar way. The regressions are generally all consistent with the proposition that American speech forms tend to be most similar to those immediately surrounding London and (mildly) progressively less similar in more distant regions, with the relation apparently stronger for Massachusetts speakers than for southern speakers. However, the regional affiliation of English informants is consistently more important in accounting for affinities with American informants than is distance from the metropolis.

3.10 Conclusions

In summary, the application of a variety of quantitative techniques to patterns of usage by twentieth-century English and American informants appears to provide a variety of insights into the nature of southern English and American speech. The analysis reveals a tremendous amount of diversity among southern English informants, distinguishes six more-or-less distinct southern English dialect regions, similar in geographic distribution to regions delineated in previous studies, and shows that the variants found among American informants were nearly all found among at least some informants in southern England. That finding appears to indicate widespread preservation of English variants and relatively little phonetic innovation in America – at least in the sense of creation of entirely distinct phonemes. As gauged by several different measures, the American varieties of English analyzed here appear to be quite comfortably placed in the family of southern English dialects, at least in terms of their phonetic characteristics, and American varieties appear to differ from their English counterparts primarily in composition. At the same time, the overall diversity within and among American regions is considerably less than that within and among the English regions, suggesting extensive – but different – leveling processes and the extinction of many – but different – English variants in each American region.

Variants found in American regions are typically more likely to be found in the southeastern regions of England and particularly in the southeast closest to London. The similarity has structural elements, including important phonetic mergers, raising of low front vowels, and retraction and rounding of back ones. However, the analysis also reveals different leveling processes in different American regions: variants found in Massachusetts, taken as a whole, tend to be considerably different from those observed in the two southern American regions, while the southern regions tend to be much more similar to each other in

usage. Informants from Massachusetts consistently show substantially greater similarity with East Midlands informants than with those of southwestern England. In contrast, variants common among southern American informants but not found in Massachusetts often appear to be more similar to those found in southwestern England – these similarities all despite the passage of centuries since the earliest colonization, and consistent with Cleanth Brooks' observation more than sixty years ago that southern American English was "strongly colored" by that of southwestern England.[19] The secondary influence of the East Midlands on Massachusetts and of the English Southwest on the American south also involves identifiable structural elements: with rhoticity, palatalization, and certain shifts in back vowels present in the latter regions and absent in the former ones.

American phonetic speech patterns thus appear to be largely amalgams of southern English variants, with a dominant influence from the regions closest to the capital but with significant East Midlands influence in the New England and greater southwestern influence in Virginia. Except for the absence of clear East Anglian influence on the speech of Massachusetts, the results are largely consistent with the historical record of the regional migrations from seventeenth- and eighteenth-century Britain to North America, which suggests that the Puritan migration to Massachusetts drew largely from the eastern counties of England, while the migration to the Tidewater region drew mainly from the metropolitan center around London and from the southwest.

The relative uniformity of American speech may stem in part from the dominance of immigrants from southeastern England. Perhaps a third of the British immigrants to America came from near London while other regions tended to contribute smaller shares to the total immigration. Of the 155,000 English immigrants who settled in the mainland North American colonies in the seventeenth century, the bulk were indentured servants who sailed from London and came from the Thames valley. Despite changes in the regional patterns of emigration during the eighteenth century, the bulk of English settlers continued to come from the southeast.[20] It seems very likely that those migrants formed a large enough portion of the early immigrant population that their speech forms tended to dominate in the development of distinctive American colonial varieties, contributing to the leveling process. As a result, metropolitan (or near-metropolitan) variants would likely have been spoken by a large share of the early settlers, or possibly accorded somewhat greater prestige, or both.

It is important to take into account the fact that during the period of colonization, internal migration was bringing an enormous number of people from

[19] See Brooks (1935), p. 73.

[20] For extensive discussion of seventeenth- and eighteenth-century immigration patterns see Bailyn (1986) and Fischer (1989).

all over Britain to London, and that that process could reasonably be expected to have influenced rural speech in the region of London. Similarities between American and Southeastern English speech may therefore simply indicate that a process parallel to the leveling that occurred in the colonies also occurred in the Southeast, resulting in relatively large similarities between those regions. As noted previously, however, informants from rather near London do not tend to cluster into a distinctive metropolitan group, or even to cluster strongly with a single region. Rather, they seem to have stronger affinities with regions not clearly related to London at all than they do with each other. To be sure, Southeastern informants tend to show somewhat greater affinity with those from the East Midlands, but that comparative similarity of East Midlands and Southeastern speech is of very long standing and is thought to be related to population movements preceding the early modern era.[21] It therefore seems unlikely that the relatively high degree of similarity between American speech and that of the English Southeast is due primarily to leveling in the vicinity of London.[22]

Those conclusions may help explain why eighteenth-century English visitors noticed little variation in American regional speech forms. If they were familiar with southeastern English rural speech, American speech probably struck them not only as similar but as similar in its variation. As London speech and an English standard each became increasingly distinct from the surrounding, increasingly less prestigious rural dialects, and as American speech forms evolved independently, nineteenth-century English visitors to America most likely would have noted both the uniformity of American speech and the difference between American speech and proper English, but would not necessarily have noted a distinct similarity of American speech to any particular English dialect.

The patterns of variation – increasing numbers of variants in regions with more informants and with longer-settled populations; high diversity in the home country coupled with extensive but varying patterns of leveling in the colonies – are reminiscent of the species-age-area relationships found in population biology and the effects of evolutionary bottlenecks found in the analysis of population genetics. A process of leveling – analogous to the loss of species during reduction in habitat – appears to have reduced the population of variants during the colonial settlement of North America, leaving American speech patterns relatively uniform, though with differences that may be traced back

[21] Anderson (1987), p. 3, argues that the "evidence points to a transfer of population from the central Midlands to the South East such as is known to have occurred in the medieval period."

[22] Despite the massive internal migrations, London itself was still a fairly small city in the seventeenth century, mainly because extremely high mortality rates largely kept pace with the rate of migration.

to differences in the founding populations. Thus the similarity between eastern and western Virginia speech forms suggests that the dominant influence on the development of speech patterns in the American south came from the regions of earliest settlement, providing support for Mufwene's Founder Principle.[23] On the whole, the results are consistent with Mufwene's model of competition and selection of linguistic features, in which "units and principles selected from different varieties ... are restructured into a new system."[24] They are also consistent with Kretzschmar's contention that

> linguistic features were retained from the habits of individual speakers, but not whole linguistic systems from constituent immigrant languages or dialects. The default condition for English in the colonies in the seventeenth centuries was no "London standard" (whatever that could have meant given the great population mobility of the time), but instead a pool of linguistic features collected from a radically mixed settlement population.[25]

A largely southeastern English origin for American speech is also consistent with the recognition of a largely southeastern English origin for other forms of colonial English. On the basis of the present analysis, it seems very likely that processes quite similar to those that produced new varieties of English in Australia and New Zealand in the nineteenth century produced American English forms during the seventeenth and early eighteenth centuries.

[23] See Mufwene (1996).
[24] See Mufwene (2002).
[25] See Kretzschmar (2002).